Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Online Network Revenue Management using Thompson Sampling

Kris Johnson Ferreira
Harvard Business School, Boston, MA 02163, kferreira@hbs.edu

David Simchi-Levi
Institute for Data, Systems, and Society, Department of Civil and Environmental Engineering, and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, dslevi@mit.edu

He Wang
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, he.wang@isye.gatech.edu

We consider a price-based network revenue management problem where a retailer aims to maximize revenue from multiple products with limited inventory over a finite selling season. As common in practice, we assume the demand function contains unknown parameters, which must be learned from sales data. In the presence of these unknown demand parameters, the retailer faces a tradeoff commonly referred to as the *exploration-exploitation tradeoff*. Towards the beginning of the selling season, the retailer may offer several different prices to try to learn demand at each price ("exploration" objective). Over time, the retailer can use this knowledge to set a price that maximizes revenue throughout the remainder of the selling season ("exploitation" objective). We propose a class of dynamic pricing algorithms that builds upon the simple yet powerful machine learning technique known as *Thompson sampling* to address the challenge of balancing the exploration-exploitation tradeoff under the presence of inventory constraints. Our algorithms prove to have both strong theoretical performance guarantees as well as promising numerical performance results when compared to other algorithms developed for similar settings. Moreover, we show how our algorithms can be extended for use in general multi-armed bandit problems with resource constraints, with applications in other revenue management settings and beyond.

Key words: revenue management, dynamic pricing, demand learning, multi-armed bandit, Thompson sampling, machine learning

1. Introduction

In this paper, we consider a price-based revenue management problem common to many retail settings: given an initial inventory of products and finite selling season, a retailer must choose

prices to maximize revenue over the course of the season. Inventory decisions are fixed prior to the selling season, and inventory cannot be replenished throughout the season. The retailer has the ability to observe consumer demand in real-time and can dynamically adjust the price at negligible cost. We refer the readers to Talluri and van Ryzin (2005) and Özer and Phillips (2012) for many applications of this revenue management problem. More generally, our work focuses on the *network revenue management* problem (Gallego and Van Ryzin 1997), where the retailer must price several unique products, each of which may consume common resources with limited inventory.

The price-based network revenue management problem has been well-studied in the academic literature, often under the additional assumption that the mean demand rate (*i.e.*, expected demand per unit time) associated with each price is known to the retailer prior to the selling season. In practice, many retailers do not know the mean demand rates for each price; thus, we focus on the network revenue management problem with unknown demand. Given unknown mean demand rates, the retailer faces a tradeoff commonly referred to as the *exploration-exploitation tradeoff*. Towards the beginning of the selling season, the retailer may offer several different prices to try to learn and estimate the mean demand rate at each price ("exploration" objective). Over time, the retailer can use these mean demand rate estimates to set a price that maximizes revenue throughout the remainder of the selling season ("exploitation" objective). In our setting, the retailer is constrained by limited inventory and thus faces an additional tradeoff. Specifically, pursuing the exploration objective comes at the cost of diminishing valuable inventory. Simply put, if inventory is depleted while exploring different prices, there is no inventory left to exploit the knowledge gained.

We will refer to the network revenue management setting with unknown mean demand rates as the *online network revenue management* problem, where "online" refers to two characteristics. First, "online" refers to the retailer's ability to observe and learn demand as it occurs throughout the selling season – in an online fashion – allowing the retailer to consider the exploration-exploitation tradeoff. Second, "online" can also refer to the online retail industry, since many online retailers face the challenge of pricing many products in the presence of demand uncertainty and short product life cycles; furthermore, many online retailers are able to observe and learn demand in real time and can easily adjust prices dynamically. The online retail industry has experienced approximately 10% annual growth over the last 5 years in the United States, reaching nearly $300B in revenue in 2015 (excluding online sales of brick-and-mortar stores); see industry report by Lerman (2014).

Motivated by this large and growing industry, we develop a class of algorithms for the online network revenue management problem. Our algorithms adapt a simple yet powerful machine learning

technique known as *Thompson sampling* to address the challenge of balancing the exploration-exploitation tradeoff under the presence of inventory constraints. In the following section, we outline the academic literature that has addressed similar revenue management challenges and describe how our work fits in this space. Then in Section 1.2 we provide an overview of the main contribution of our paper to this body of literature and to practice.

1.1. Literature Review

Due to the increased availability of real-time demand data, there is a vast literature on dynamic pricing problems with a demand learning approach. Review papers by Aviv and Vulcano (2012) and den Boer (2015) provide up-to-date surveys of this area. Our review below on dynamic pricing with demand learning is mostly focused on existing literature that considers *inventory constraints*.

As described earlier, the key challenge in dynamic pricing with demand learning is to address the exploration-exploitation tradeoff, where the retailer's ability to learn demand is tied to the actions the retailer takes (*e.g.*, the prices the retailer offers). Several approaches have been proposed in the literature to address the exploration-exploitation tradeoff in the constrained inventory setting.

One approach is to separate the selling season (T periods) into a disjoint exploration phase (say, from period 1 to τ) and exploitation phase (from period $\tau+1$ to T) (see, *e.g.*, Besbes and Zeevi (2009, 2012)). During the exploration phase, each price is offered for a pre-determined number of times. At the end of period τ, the retailer uses purchasing data from the first τ periods to estimate the mean demand rate for each price. These estimates are then used ("exploited") to maximize revenue during periods $\tau+1$ to T. One drawback of this strategy is that it does not use purchasing data after period τ to continuously refine its estimates of the mean demand rates for each price. Furthermore, when there is very limited inventory, this approach is susceptible to running out of inventory during the exploration phase, before any demand learning can be exploited. We note that Besbes and Zeevi (2012) considers a similar online network revenue management setting as we do, and in Section 3.2, we will compare the performance of their algorithm with ours via numerical experiments.

A second approach is to model the online network revenue management problem as a multi-armed bandit problem and use a popular method known as the upper confidence bound (UCB) algorithm (Auer et al. 2002) to dictate pricing decisions in each period. The multi-armed bandit (MAB) problem is often used to model the exploration-exploitation tradeoff in the dynamic learning and pricing model *without* limited inventory constraints since it can be immediately applied to such a setting; see Bubeck and Cesa-Bianchi (2012) for an overview of this problem. The UCB algorithm creates a confidence interval for each unknown mean demand rate using purchase data and then selects a price that maximizes revenue among all parameter values in the confidence set. For the

purpose of exploration, the UCB algorithm favors prices that have not been offered many times since they are associated with a larger confidence interval. The presence of operational constraints such as limited inventory cannot be directly modeled in the standard MAB problem; Badanidiyuru et al. (2013) thus builds upon the MAB problem and adapts the UCB algorithm to a setting with inventory constraints. In Section 3.2, we will compare the performance of our algorithms to the algorithm in Badanidiyuru et al. (2013) via numerical experiments.

There are several other methods developed for revenue management problems with unknown demand in limited inventory settings; the models in the following papers are different than the model in our setting and thus we only compare our algorithms to those presented in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013). Araman and Caldentey (2009) and Farias and Van Roy (2010) use dynamic programming to study settings with unknown market size but with known customer willingness-to-pay function. Chen et al. (2014) considers a strategy that separates exploration and exploitation phases, while using self-adjusting heuristics in the exploitation phase. Wang et al. (2014) proposes a continuously learning-and-optimization algorithm for a single product and continuous price setting. Lastly, Jasin (2015) studies a quantity-based revenue management model with unknown parameters; in a quantity-based model, the retailer observes all customer arrivals and either accepts or rejects their purchase requests, so the retailer is not faced with the same type of exploration-exploitation tradeoff as in the price-based model.

It is worth noting that several of the papers reviewed above consider only the continuous-price setting (Araman and Caldentey (2009), Besbes and Zeevi (2009), Farias and Van Roy (2010), Wang et al. (2014), and Chen et al. (2014)), whereas our work primarily considers the discrete-price setting with an extension to the continuous-price, linear demand setting presented in Section 4.1. We choose to focus primarily on the discrete-price setting because discrete price sets are widely used in practice (Talluri and van Ryzin (2005)). A key distinction between discrete vs. continuous price sets in demand learning and dynamic pricing arises from the structure of their respective revenue optimization problems. If the price set is discrete and the retailer faces inventory constraints, there may not exist any single price that maximizes revenue asymptotically as the number of periods T increases; the retailer must maximize over *distributions of prices*. In contrast, if the price set is continuous and the demand function satisfies certain regularity conditions (Gallego and Van Ryzin 1994), there always exists a *single* price that is asymptotically optimal, regardless of the presence of inventory constraints. Demand learning and dynamic pricing algorithms developed for continuous price sets rely on the fact that a single optimal price exists, and therefore it is difficult to immediately extend them to discrete price sets in the presence of inventory constraints; thus, we cannot compare the performance of our algorithms to those developed for the continuous-price setting.

Our approach is most closely related to the second approach summarized above and used in Badanidiyuru et al. (2013), in that we also model the online network revenue management problem as a multi-armed bandit problem with inventory constraints. However, rather than using the UCB algorithm as the backbone of our algorithms, we use the powerful machine learning algorithm known as *Thompson sampling* as a key building block of the algorithms that we develop for the online network revenue management problem.

Thompson sampling. In one of the earliest papers on the multi-armed bandit problem, Thompson (1933) proposed a randomized Bayesian algorithm, which was later referred to as the *Thompson sampling* algorithm. The basic idea of Thompson sampling is that at each time period, random numbers are sampled according to the posterior distributions of the reward for each action, and then the action with the highest sampled reward is chosen; a formal description of the algorithm can be found in the Appendix. Note that in a revenue management setting, each "action" or "arm" is a price, and "reward" refers to the revenue earned by offering that price. Thus in the original Thompson sampling algorithm – in the absence of inventory constraints – random numbers are sampled according to the posterior distributions of the mean demand rates for each price, and the price with the highest sampled revenue (*i.e.*, price times sampled demand) is offered. Thompson sampling is also known as *probability matching* since the probability of an arm being chosen matches the posterior probability that this arm has the highest expected reward.

This randomized Bayesian approach is in contrast to the more traditional Bayesian "greedy" approach, where instead of sampling from the posterior probability distributions, the expected value of each posterior distribution is used to evaluate the reward of each arm (expected revenue for each price offered). Such a greedy approach makes decisions solely with the exploitation goal in mind by choosing the price that is believed to be optimal in the current period; this approach does not actively explore by deviating from greedy prices, and therefore might get stuck with a suboptimal price forever. Harrison et al. (2012) illustrates the potential pitfalls of such a greedy Bayesian approach, and shows the necessity to deviate from greedy prices in order to get sufficient exploration. Thompson sampling satisfies the exploration objective by using random samples that deviate from the greedy optimal solution.

Thompson sampling enjoys similar theoretical performance guarantees to those achieved by other popular multi-armed bandit algorithms such as the UCB algorithm (Kaufmann et al. 2012, Agrawal and Goyal 2013) and often better empirical performance (Chapelle and Li 2011). In addition, the Thompson sampling algorithm has been adapted to various multi-armed bandit settings by Russo and Van Roy (2014). In our work, we adapt Thompson sampling to the network revenue management setting where inventory is constrained, thus bridging the gap between a popular machine learning technique for the exploration-exploitation tradeoff and a common revenue management challenge.

1.2. Overview of Main Contribution

The main contribution of our work is the design and development of a new class of algorithms for the online network revenue management problem: this class of algorithms extends the powerful machine learning technique known as Thompson sampling to address the challenge of balancing the exploration-exploitation tradeoff *under the presence of inventory constraints*. We first consider a model with discrete price sets in Section 2.1, as this is a common constraint that is self-imposed by many retailers in practice. In Section 2.2, we present our first algorithm which adapts Thompson sampling by adding a linear programming (LP) subroutine to incorporate inventory constraints. In Section 2.3, we present our second algorithm that builds upon our first; specifically, in each period, we modify the LP subroutine to further account for the purchases made to date. Both of our algorithms contain two simple steps in each iteration: sampling from a posterior distribution and solving a linear program. As a result, the algorithms are easy to implement in practice.

To highlight the importance of our main contribution, Section 3 provides both a theoretical and numerical performance analysis of both of our algorithms. In Section 3.1, we show the proposed algorithms have strong theoretical performance guarantees. We measure the algorithms' performance by *regret*, *i.e.*, the difference in expected revenue obtained by our algorithms compared to the expected revenue of the ideal case where the mean demand rates are known at the beginning of the selling season. More specifically, since Thompson sampling is defined in a Bayesian setting, our measurement is focused on *Bayesian regret* (defined in Section 3.1.1). We show that our proposed algorithms have a Bayesian regret of $O(\sqrt{TK \log K})$, where T is the length of the selling season and K is the number of feasible price vectors. Since this bound depends on T by $O(\sqrt{T})$, our bound matches the best possible prior-free lower bound for Bayesian regret, $\Omega(\sqrt{T})$ (Bubeck and Cesa-Bianchi 2012). Our proof heavily builds on the techniques of Bubeck and Liu (2013) and Russo and Van Roy (2014) for analyzing Thompson sampling. As a technical contribution, we show how the tools of Bubeck and Liu (2013) and Russo and Van Roy (2014) can be modified to analyze lost sales, a quantity of nonlinear form; their tools were originally developed for analyzing rewards in multi-armed bandit problems, which has a linear and additive form. In Section 3.2, we present numerical experiments which show that our algorithms have significantly better empirical performance than the algorithms developed for similar settings by Badanidiyuru et al. (2013) and Besbes and Zeevi (2012).

Finally, in Section 4, we broaden our main contribution by showing how our algorithms can be adapted to address various other revenue management and operations management challenges. Specifically, we consider three extensions: 1) continuous price sets with a linear demand function; 2) dynamic pricing with contextual information; 3) multi-armed bandits with general resource

constraints. Using the general recipe of combining Thompson sampling with an LP subroutine, we show that our algorithms can be naturally extended to these problems and have an $\tilde{O}(\sqrt{T})$ regret bound (omitting log factors) in all three settings.

2. Discrete Price Thompson Sampling with Limited Inventory

We start by focusing on the case where the set of possible prices that the retailer can offer is discrete and finite as this is a common constraint that is self-imposed by many retailers in practice (Talluri and van Ryzin 2005). We first introduce our model formulation in Section 2.1, and then we propose two dynamic pricing algorithms based on Thompson sampling for this model setting in Sections 2.2 and 2.3. Both algorithms incorporate inventory constraints into the original Thompson sampling algorithm, which is included in the Appendix for reference. In Section 4 we provide extensions of our algorithms to the continuous price setting as well as other operations management settings.

2.1. Discrete Price Model

We consider a retailer who sells N products, indexed by $i \in [N]$, over a finite selling season. (Throughout the paper, we denote by $[x]$ the set $\{1, 2, \ldots, x\}$.) These products consume M resources, indexed by $j \in [M]$. Specifically, we assume that one unit of product i consumes a_{ij} units of resource j, where a_{ij} is a fixed constant. The selling season is divided into T periods. There are I_j units of initial inventory for each resource $j \in [M]$, and there is no replenishment during the selling season. We define $I_j(t)$ as the inventory at the end of period t, and we denote $I_j(0) = I_j$. In each period $t \in [T]$, the following sequence of events occurs:

1. The retailer offers a price for each product from a finite set of admissible price vectors. We denote this set by $\{p_1, p_2, \ldots, p_K\}$, where p_k $(\forall k \in [K])$ is a vector of length N specifying the price of each product. More specifically, we have $p_k = (p_{1k}, \ldots, p_{Nk})$, where p_{ik} is the price of product i, for all $i \in [N]$. Following the tradition in dynamic pricing literature, we also assume that there is a "shut-off" price p_∞ such that the demand for any product under this price is zero with probability one. We denote by $P(t) = (P_1(t), \ldots, P_N(t))$ the prices chosen by the retailer in this period, and require that $P(t) \in \{p_1, p_2, \ldots, p_K, p_\infty\}$.

2. Customers then observe the prices chosen by the retailer and make purchase decisions. We denote by $D(t) = (D_1(t), \ldots, D_N(t))$ the demand of each product at period t. We assume that given $P(t) = p_k$, the demand $D(t)$ is sampled from a probability distribution on $\mathbb{R}_+^N$ with joint cumulative distribution function (CDF) $F(x_1, \ldots, x_N; p_k, \theta)$, indexed by a parameter (or a vector of parameters) θ that takes values in the parameter space $\Theta \subset \mathbb{R}^l$. The distribution is assumed to be subexponential; note that many commonly used demand distributions such as normal, Poisson, exponential and all bounded distributions belong to the

family of subexponential distributions. We also assume that $D(t)$ is independent of the history $\mathcal{H}_{t-1} = (P(1), D(1), \ldots, P(t-1), D(t-1))$ given $P(t)$.

Depending on whether there is sufficient inventory, one of the following events happens:

(a) If there is enough inventory to satisfy all demand, the retailer receives an amount of revenue equal to $\sum_{i=1}^{N} D_i(t)P_i(t)$, and the inventory level of each resource $j \in [M]$ diminishes by the amount of each resource used such that $I_j(t) = I_j(t-1) - \sum_{i=1}^{N} D_i(t)a_{ij}$.

(b) If there is not enough inventory to satisfy all demand, the demand is partially satisfied and the rest of demand is lost. Let $\tilde{D}_i(t)$ be the demand satisfied for product i. We require $\tilde{D}_i(t)$ to satisfy three conditions: $0 \le \tilde{D}_i(t) \le D_i(t), \forall i \in [N]$; the inventory level for each resource at the end of this period is nonnegative: $I_j(t) = I_j(t-1) - \sum_{i=1}^{N} \tilde{D}_i(t)a_{ij} \ge 0, \forall j \in [M]$; there exists at least one resource $j' \in [M]$ whose inventory level is zero at the end of this period, *i.e.* $I_{j'}(t) = 0$. Besides these natural conditions, we do not require any additional assumption on how demand is specifically fulfilled. The retailer then receives an amount of revenue equal to $\sum_{i=1}^{N} \tilde{D}_i(t)P_i(t)$ in this period.

We assume that the demand parameter θ is fixed but *unknown* to the retailer at the beginning of the season, and the retailer must learn the true value of θ from demand data. That is, in each period $t \in [T]$, the price vector $P(t)$ can only be chosen based on the observed history $\mathcal{H}_{t-1}$, but cannot depend on the unknown value θ or any event in the future. The retailer's objective is to maximize expected revenue over the course of the selling season given the prior distribution on θ.

We use a parametric Bayesian approach in our model, where the retailer has a *known* prior distribution of $\theta \in \Theta$ at the beginning of the selling season. However, our model allows the retailer to choose an arbitrary prior. In particular, the retailer can assume an arbitrary parametric form of the demand CDF, given by $F(x_1, \ldots, x_N; p_k, \theta)$. This joint CDF parametrized by θ can parsimoniously model the correlation of demand among products. For example, the retailer may specify products' joint demand distribution based on some discrete choice model such as the multinomial logit model, where θ is the unknown parameter in the multinomial logit function. Another benefit of the Bayesian approach is that the retailer may choose a prior distribution over θ such that demand is correlated for different prices, enabling the retailer to learn demand for all prices, not just the offered price. For example, with a single product, the retailer may assume that demand follows a Poisson distribution in each period, and the mean demand is a linear function in price. The retailer may assume there are two unknown parameters $\theta = (\theta_0, \theta_1)$ such that the mean demand under price p is given by $\theta_0 - \theta_1 p$. Therefore, the Poisson CDF as a function of price $p_k \in \{p_1, p_2, \ldots, p_K\}$ and θ is given by

$$F(x; p_k, \theta) = \sum_{i=0}^{x} \frac{(\theta_0 - \theta_1 p_k)^i}{i!} e^{\theta_1 p_k - \theta_0}.$$

When the retailer observes a realized demand instance under the offered price $p_k \in \{p_1, p_2, \ldots, p_K\}$, it obtains some information about parameters θ_0 and θ_1, which enables the retailer to learn demand not only for the offered price, but also for prices that are not offered.

Relationship to the Multi-Armed Bandit Problem The model formulated above is a generalization of the multi-armed bandit (MAB) problem that has been extensively studied in the statistics and operations research literature – where each price is an "arm" and revenue is the "reward" – except for two main deviations. First, our formulation allows for the network revenue management setting (Gallego and Van Ryzin (1997)) where multiple products consuming common resources are sold. Second, there are inventory constraints present in our setting, whereas there are no such constraints in the MAB model.

We note that the presence of inventory constraints significantly complicates the problem, even for the special case of a single product. In the MAB setting, if mean revenue associated with each price vector is known, the optimal strategy is to choose a price vector with the highest mean revenue. But in the presence of limited inventory, a mixed strategy that chooses multiple price vectors over the selling season may achieve significantly higher revenue than any single price strategy. Therefore, a good pricing strategy should converge not to a single price, but to a distribution of (possibly) multiple prices. Another challenging task in the analysis is to estimate the time when the inventory of each resource runs out, which is itself a random variable depending on the pricing policy used by the retailer. Such estimation is necessary for computing the retailer's expected revenue. This is in contrast to classical MAB problems for which the process always ends at a fixed period.

Our model is also closely related to the models studied in Badanidiyuru et al. (2013) and Besbes and Zeevi (2012). Badanidiyuru et al. (2013) considers a multi-armed bandit problem with global resource constraints. We will discuss this problem and extend our algorithms to this setting in Section 4.3. Besbes and Zeevi (2012) studies a similar network revenue management model with continuous time and unknown demand, considering both discrete and continuous price sets. Our model can incorporate their setting by discretizing time, and we will discuss the extension to continuous price sets in Section 4.1.

2.2. Thompson Sampling with Fixed Inventory Constraints

In this section, we propose our first Thompson sampling based algorithm for the discrete price model described in Section 2.1. For each resource $j \in [M]$, we define a fixed constant $c_j := I_j/T$. Given any demand parameter $\rho \in \Theta$, we define the mean demand under ρ as the expectation associated with CDF $F(x_1, \ldots, x_N; p_k, \rho)$ for each product $i \in [N]$ and price vector $k \in [K]$. We denote by $d = \{d_{ik}\}_{i\in[N],k\in[K]}$ the mean demand under the *true* model parameter θ.

We present our Thompson Sampling with Fixed Inventory Constraints algorithm (TS-fixed for short) in Algorithm 1. Here, "TS" stands for Thompson sampling, while "fixed" refers to the fact that we use fixed constants c_j for all time periods as opposed to updating c_j over the selling season as inventory is depleted; this latter idea is incorporated into the algorithm we present in Section 2.3.

Algorithm 1 Thompson Sampling with Fixed Inventory Constraints (TS-fixed)

Repeat the following steps for all periods $t = 1, ..., T$:

1. *Sample Demand*: Sample a random parameter $\theta(t) \in \Theta$ according to the posterior distribution of θ given history $\mathcal{H}_{t-1}$. Let the mean demand under $\theta(t)$ be $d(t) = \{d_{ik}(t)\}_{i\in[N],k\in[K]}$.

2. *Optimize Prices given Sampled Demand*: Solve the following linear program, denoted by $\mathsf{LP}(d(t))$:

$$\begin{aligned}
\mathsf{LP}(d(t)): \quad \max_{x} \ & \sum_{k=1}^{K}\Big(\sum_{i=1}^{N} p_{ik} d_{ik}(t)\Big)x_k \\
\text{subject to } & \sum_{k=1}^{K}\Big(\sum_{i=1}^{N} a_{ij} d_{ik}(t)\Big)x_k \le c_j, \ \forall j \in [M] \\
& \sum_{k=1}^{K} x_k \le 1 \\
& x_k \ge 0, \ k \in [K].
\end{aligned}$$

Let $x(t) = (x_1(t), \ldots, x_K(t))$ be the optimal solution to $\mathsf{LP}(d(t))$.

3. *Offer Price*: Offer price vector $P(t) = p_k$ with probability $x_k(t)$, and choose $P(t) = p_\infty$ with probability $1 - \sum_{k=1}^{K} x_k(t)$.

4. *Update Estimate of Parameter*: Observe demand $D(t)$. Update the history $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{P(t), D(t)\}$ and the posterior distribution of θ given $\mathcal{H}_t$.

Steps 1 and 4 are based on the Thompson sampling algorithm for the classical multi-armed bandit setting, whereas steps 2 and 3 are added to incorporate inventory constraints. In step 1 of the algorithm, we randomly sample parameter $\theta(t)$ according to the posterior distribution of unknown demand parameter θ. This step is motivated by the original Thompson sampling algorithm for the classical multi-armed bandit problem. A novel idea of the Thompson sampling algorithm is to use *random sampling from the posterior distribution to balance the exploration-exploitation tradeoff.* To be more precise, let us consider an example when there is unlimited inventory. Without loss of generality, let us assume that price vector p_1 has the highest expected revenue under the posterior distribution in the current period. If the retailer acts greedily (*i.e.* focusing only on the exploitation

objective), it would maximize the expected revenue in this period by choosing p_1 with probability one. However, there is no guarantee that p_1 is indeed the optimal price under the true demand. In Thompson sampling, the retailer balances the exploration-exploitation tradeoff by using demand values that are randomly sampled, which means that there is a positive probability that the retailer will choose a price vector other than p_1, thus achieving the exploration objective. Guaranteeing positive probability to pursue each objective - exploration and exploitation - is essential to discover the true demand parameter over time (cf. Harrison et al. 2012).

The algorithm differs from ordinary Thompson sampling in steps 2 and 3. In step 2, the retailer solves a linear program, $\mathsf{LP}(d(t))$, which identifies the optimal mixed price strategy that maximizes expected revenue given the sampled parameters. The first constraint specifies that the average resource consumption in this time period cannot exceed c_j, the average inventory available per period. The second constraint specifies that the sum of probabilities of choosing a price vector cannot exceed one. In step 3, the retailer randomly offers one of the K price vectors (or p_∞) according to probabilities specified by the optimal solution of $\mathsf{LP}(d(t))$. Finally, in step 4, the algorithm updates the posterior distribution of θ given $\mathcal{H}_t$. Such Bayesian updating is a simple and powerful tool to update belief probabilities as more information – customer purchase decisions in our case – becomes available. By employing Bayesian updating in step 4, we are ensured that as any price vector p_k is offered more and more times, the sampled mean demand associated with p_k for each product i becomes more and more centered around the true mean demand, d_{ik} (cf. Freedman 1963).

We note that the LP defined in step 2 is closely related to the LP used by Gallego and Van Ryzin (1997), where they consider a network revenue management problem in the case of known demand. Their pricing algorithm is essentially a special case of Algorithm 1 where they solve $\mathsf{LP}(d)$, i.e, $\mathsf{LP}(d(t))$ with $d(t) = d$, in every time period. Moreover, they show that the optimal value of $\mathsf{LP}(d)$ is an upper bound on the expected optimal revenue that can be achieved in such a network revenue management setting; in Section 3.1.1 we present this upper bound and discuss the similarities between the two linear programs.

Next we illustrate the application of our TS-fixed algorithm by providing two concrete examples. For simplicity, in both examples we assume that the prior distribution of demand for different prices are independent; however, the definition of TS-fixed and the theoretical results in Section 3.1 are quite general and allow the prior distribution to be *arbitrarily correlated* for different prices. As mentioned earlier, this enables the retailer to learn the mean demand not only for the offered price, but also for prices that are not offered.

Example 1: Bernoulli Demand with Independent Uniform Prior We assume that for all prices, the demand for each product is Bernoulli distributed. In this case, the unknown parameter θ is just the mean demand of each product. We use a beta posterior distribution for each θ because it is conjugate to the Bernoulli distribution. We assume that the prior distribution of mean demand d_{ik} is uniform in $[0,1]$ (which is equivalent to a Beta$(1,1)$ distribution) and is independent for all $i \in [N]$ and $k \in [K]$.

In this example, the posterior distribution is very simple to calculate. Let $N_k(t-1)$ be the number of time periods that the retailer has offered price p_k in the first $t-1$ periods, and let $W_{ik}(t-1)$ be the number of periods that product i is purchased under price p_k during these periods. In step 1 of TS-fixed, the posterior distribution of d_{ik} is Beta$(W_{ik}(t-1)+1, N_k(t-1)-W_{ik}(t-1)+1)$, so we sample $d_{ik}(t)$ independently from a Beta$(W_{ik}(t-1)+1, N_k(t-1)-W_{ik}(t-1)+1)$ distribution for each price k and each product i.

In steps 2 and 3, LP$(d(t))$ is solved and a price vector $p_{k'}$ is chosen; then the customer demand $D_i(t)$ is revealed to the retailer. In step 4, we then update $N_{k'}(t) \leftarrow N_{k'}(t-1)+1$, $W_{ik'}(t) \leftarrow W_{ik'}(t-1)+D_i(t)$ for all $i \in [N]$. The posterior distributions associated with the $K-1$ unchosen price vectors ($k \neq k'$) are not changed.

Example 2: Poisson Demand with Independent Exponential Prior We now consider another example where demand for each product follows a Poisson distribution. Like the previous example, the unknown parameter θ is just the mean demand of each product. We use a gamma posterior distribution for each θ because it is conjugate to the Poisson distribution. We assume that the prior distribution of mean demand d_{ik} is exponential with CDF $f(x) = e^{-x}$ (which is equivalent to a Gamma$(1,1)$ distribution) and is independent for all $i \in [N]$ and $k \in [K]$.

The posterior distribution is also simple to calculate in this case. Let $N_k(t-1)$ be the number of time periods that the retailer has offered price vector p_k in the first $t-1$ periods, and let $W_{ik}(t-1)$ be the total demand for product i during these periods. In step 1 of TS-fixed, the posterior distribution of d_{ik} is Gamma$(W_{ik}(t-1)+1, N_k(t-1)+1)$, so we sample $d_{ik}(t)$ independently from a Gamma$(W_{ik}(t-1)+1, N_k(t-1)+1)$ distribution for each price k and each product i.

In steps 2 and 3, LP$(d(t))$ is solved and the price vector $P(t) = p_{k'}$ for some $k' \in [K]$ is chosen; then the customer demand $D_i(t)$ is revealed to the retailer. In step 4, we then update $N_{k'}(t) \leftarrow N_{k'}(t-1)+1$, $W_{ik'}(t) \leftarrow W_{ik'}(t-1)+D_i(t)$ for all $i \in [N]$. The posterior distributions associated with the $K-1$ unchosen price vectors ($k \neq k'$) are not changed.

2.3. Thompson Sampling with Inventory Constraint Updating

In this section, we propose our second Thompson sampling based algorithm for the discrete price model described in Section 2.1. In TS-fixed, we use fixed inventory constants c_j in every period. Alternatively, we can update c_j over the selling season as inventory is depleted, thereby incorporating real time inventory information into the algorithm.

In particular, we recall that $I_j(t)$ is the inventory level of resource j at the end of period t. Define $c_j(t) = I_j(t-1)/(T-t+1)$ as the average inventory for resource j available from period t to period T. We then replace constants c_j with $c_j(t)$ in LP$(d(t))$ in step 2 of TS-fixed, which gives us the Thompson Sampling with Inventory Constraint Updating algorithm (TS-update for short) shown in Algorithm 2. The term "update" refers to the fact that in every iteration, the algorithm updates inventory constants $c_j(t)$ in LP$(d(t))$ to incorporate real time inventory information.

Algorithm 2 Thompson Sampling with Inventory Constraint Updating (TS-update)

Repeat the following steps for all periods $t = 1, ..., T$:

1. *Sample Demand*: Sample a random parameter $\theta(t) \in \Theta$ according to the posterior distribution of θ given history $\mathcal{H}_{t-1}$. Let the mean demand under $\theta(t)$ be $d(t) = \{d_{ik}(t)\}_{i\in[N],k\in[K]}$.

2. *Optimize Prices given Sampled Demand*: Solve the following linear program, denoted by LP$(d(t), c(t))$:

$$\begin{aligned}
\text{LP}(d(t), c(t)): \quad \max_x \ & \sum_{k=1}^{K}\Big(\sum_{i=1}^{N} p_{ik} d_{ik}(t)\Big) x_k \\
\text{subject to } & \sum_{k=1}^{K}\Big(\sum_{i=1}^{N} a_{ij} d_{ik}(t)\Big) x_k \le c_j(t), \ \forall j \in [M] \\
& \sum_{k=1}^{K} x_k \le 1 \\
& x_k \ge 0, \ k \in [K].
\end{aligned}$$

Let $x(t) = (x_1(t), \ldots, x_K(t))$ be the optimal solution to LP$(d(t), c(t))$.

3. *Offer Price*: Offer price vector $P(t) = p_k$ with probability $x_k(t)$, and choose $P(t) = p_\infty$ with probability $1 - \sum_{k=1}^{K} x_k(t)$.

4. *Update Estimate of Parameter*: Observe demand $D(t)$. Update the history $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{P(t), D(t)\}$ and the posterior distribution of θ given $\mathcal{H}_t$.

In the revenue management literature, the idea of using updated inventory rates like $c_j(t)$ has been previously studied in various settings (Jasin and Kumar 2012, Chen and Farias 2013, Chen et al. 2014, Jasin 2015). However, to the best of our knowledge, TS-update is the first algorithm

that incorporates real time inventory updating when the retailer faces an exploration-exploitation tradeoff with its pricing decisions.[1] Although intuitively incorporating updated inventory information into the pricing algorithm should improve the performance of the algorithm, Cooper (2002) provides a counterexample where the expected revenue is reduced after the updated inventory information is included. Therefore, it is not immediately clear if TS-update would achieve higher revenue than TS-fixed. We will rigorously analyze the performance of both TS-fixed and TS-update using theoretical and numerical analysis in the next section; our numerical analysis shows that in fact there are situations where TS-update outperforms TS-fixed and vice versa.

3. Performance Analysis

To illustrate the value of incorporating inventory constraints in Thompson sampling, in Section 3.1 we prove finite-time (*i.e.* non-asymptotic) performance guarantees for TS-fixed and TS-update that match the best possible guarantees that can be achieved by any algorithm. Then in Section 3.2, we show that our algorithms outperform previously proposed algorithms for similar settings in numerical experiments.

3.1. Theoretical Results

3.1.1. Benchmark and Linear Programming Relaxation To evaluate the retailer's strategy, we compare the retailer's revenue with a benchmark where the true demand distribution is known a priori.

We define the retailer's *regret* over the selling horizon as

$$\text{Regret}(T,\theta) = \text{E}[\text{Rev}^*(T) \mid \theta] - \text{E}[\text{Rev}(T) \mid \theta],$$

where $\text{Rev}^*(T)$ is the revenue achieved by the optimal policy if the demand parameter θ is known a priori, and $\text{Rev}(T)$ is the revenue achieved by an algorithm that may not know θ. The conditional expectation is taken on random demand realizations given θ, and possibly on some external randomization used by the algorithm (*e.g.* random samples in Thompson sampling). In words, the regret is a non-negative quantity measuring the retailer's revenue loss due to not knowing the latent demand parameter.

We also define the *Bayesian regret* (also known as *Bayes risk*) by

$$\text{BayesRegret}(T) = E[\text{Regret}(T,\theta)],$$

where the expectation is taken over the prior distribution of θ. Bayesian regret is a standard metric for the performance of online Bayesian algorithms; see, *e.g.*, Rusmevichientong and Tsitsiklis (2010) and Russo and Van Roy (2014).

Because evaluating the expected optimal revenue with known demand requires solving a high dimensional dynamic programming problem, it is difficult to compute the optimal revenue exactly even for moderate problem sizes. Gallego and Van Ryzin (1997) show that the expected optimal revenue with known demand can be approximated by an upper bound. The upper bound is given by the following deterministic LP, denoted by $\mathsf{LP}(d)$:

$$\begin{aligned}\mathsf{LP}(d): \quad \max_{x} \ & \sum_{k=1}^{K}(\sum_{i=1}^{N} p_{ik}d_{ik})x_k \\ \text{subject to } & \sum_{k=1}^{K}(\sum_{i=1}^{N} a_{ij}d_{ik})x_k \le c_j,\ \forall j \in [M] \\ & \sum_{k=1}^{K} x_k \le 1 \\ & x_k \ge 0,\ \forall k \in [K].\end{aligned}$$

Problem $\mathsf{LP}(d)$ is almost identical to $\mathsf{LP}(d(t))$ used in TS-fixed, except that it uses the true mean demand d instead of sampled demand $d(t)$ from the posterior distribution. We denote the optimal value of $\mathsf{LP}(d)$ by $\mathrm{OPT}(d)$. Gallego and Van Ryzin (1997) show that

$$E[\mathrm{Rev}^*(T) \mid d] \le \mathrm{OPT}(d) \cdot T.$$

Therefore, we have

$$\mathrm{Regret}(T, d) \le \mathrm{OPT}(d) \cdot T - \mathrm{E}[\mathrm{Rev}(T) \mid d],$$

and

$$\mathrm{BayesRegret}(T) \le \mathrm{E}[\mathrm{OPT}(d)] \cdot T - \mathrm{E}[\mathrm{Rev}(T)].$$

3.1.2. Analysis of *TS-fixed* and *TS-update* Algorithms We now prove regret bounds for TS-fixed and TS-update under the realistic assumption of bounded demand. Specifically, in the following analysis, we further assume that for each product $i \in [N]$, the demand $D_i(t)$ is bounded by $D_i(t) \in [0, \bar{d}_i]$ under any price vector p_k, $\forall k \in [K]$. However, our analysis can be generalized when the demand is unbounded and follows a subexponential distribution.[2] We also define constants

$$p_{\max} := \max_{k\in[K]} \sum_{i=1}^{N} p_{ik}\bar{d}_i, \quad p^j_{\max} := \max_{i\in[N]:a_{ij}\neq 0, k\in[K]} \frac{p_{ik}}{a_{ij}},\ \forall j \in [M]$$

where $p_{\max}$ is the maximum revenue that can possibly be achieved in one period, and $p^j_{\max}$ is the maximum revenue that can possibly be achieved by adding one unit of resource j, $\forall j \in [M]$.

THEOREM 1. *The Bayesian regret of* TS-fixed *is bounded by*

$$\mathrm{BayesRegret}(T) \le \left(18p_{\max} + 37\sum_{i=1}^{N}\sum_{j=1}^{M} p^j_{\max} a_{ij}\bar{d}_i\right)\sqrt{TK\log K}.$$

THEOREM 2. *The Bayesian regret of* TS-update *is bounded by*

$$\text{BayesRegret}(T) \leq \left(18p_{\max} + 40\sum_{i=1}^{N}\sum_{j=1}^{M} p_{\max}^{j} a_{ij}\bar{d}_i\right)\sqrt{TK\log K} + p_{\max}M.$$

The results above state that the Bayesian regrets of both TS-fixed and TS-update are bounded by $O(\sqrt{TK\log K})$, where K is the number of price vectors that the retailer is allowed to use and T is the number of time periods. Moreover, the regret bounds are *prior-free* as they do not depend on the prior distribution of parameter θ; the constants in the bounds can be computed explicitly without knowing the demand distribution.

It has been shown that for a multi-armed bandit problem with reward in $[0,1]$ – a special case of our model with no inventory constraints – no algorithm can achieve a prior-free Bayesian regret smaller than $\Omega(\sqrt{KT})$ (see Theorem 3.5, Bubeck and Cesa-Bianchi (2012)). In that sense, our regret bounds are optimal with respect to T and cannot be improved by any other algorithm by more than $\sqrt{\log K}$.

The detailed proofs of Theorems 1 and 2 can be found in the E-companion. We briefly summarize the intuition behind the proofs. For both Theorems 1 and 2, we first assume an "ideal" scenario where the retailer is able to collect revenue even for the demand associated with lost sales. We show that if prices are given according to the solutions of TS-fixed or TS-update, the expected revenue achieved by the retailer is within $O(\sqrt{T})$ compared to the LP benchmark defined in Section 3.1.1. Of course, this procedure overestimates the expected revenue. In order to compute the actual revenue given constrained inventory, we should account for the amount of revenue that is associated with lost sales. For Theorem 1 (TS-fixed), we prove that the amount associated with lost sales is no more than $O(\sqrt{T})$. For Theorem 2 (TS-update), we show that the amount associated with lost sales is no more than $O(1)$.

REMARK 1. It is useful to compare the regret bounds in Theorems 1 and 2 to the regret bounds in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013), since the algorithms proposed in those papers can be applied to our model as well. However, the algorithms proposed in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013) are non-Bayesian, and they both consider the *worst case regret*, defined by

$$\max_{\theta\in\Theta} \text{Regret}(T,\theta),$$

where Θ is the set of all possible demand parameters. Besbes and Zeevi (2012) propose an algorithm with worst case regret $O(K^{5/3}T^{2/3}\sqrt{\log T})$ (Theorem 1 in their paper), while Badanidiyuru et al. (2013) provide an algorithm with worst case regret $O(\sqrt{KT\log T})$ (Theorem 4.1 in their paper). Unlike their results, our regret bounds in Theorems 1 and 2 are in terms of *Bayesian regret*, as we defined earlier in Section 3.1.1. We refer readers to Russo and Van Roy (2014) for further discussion

on Bayesian regret, and in particular, on the connection between Bayesian regret bounds and a high probability bound on Regret(T, θ).

REMARK 2. Let us remark on how the performance of TS-fixed and TS-update depends on K, the number of price vectors. Theorems 1 and 2 show that the regret bounds depend on K by $O(\sqrt{K \log K})$. Therefore, these bounds are meaningful only when K is small. Unfortunately, as the number of products increases, K may increase exponentially fast.

In practice, there are several ways to improve our algorithms' performance when K is large. First, the Thompson sampling algorithm allows any prior distribution of demand to be specified. Thus, the retailer may choose a prior distribution that is correlated for different prices. This enables the retailer to learn demand not only for the offered price, but also for prices that are not offered. We provide an example for linear demand in Section 4.1. In fact, allowing demand dependence on prices provides a major advantage over the algorithms in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013), which must learn the mean demand for each price vector independently.

Second, the retailer may have practical business constraints that it wants to impose on the price vectors. For example, many apparel retailers choose to offer the same price for different colors of the same style; each color would be a unique product since it has its own inventory and demand, but every price vector must have the same price for each of these products. Such business constraints significantly reduce the number of feasible price vectors.

REMARK 3. Note that the regret bound of TS-update is slightly worse than the regret bound of TS-fixed. Although intuition would suggest that updating inventory information in TS-update will lead to better performance than TS-fixed, this intuition is somewhat surprisingly not always true – we can find counterexamples where updating inventory information actually deteriorates the performance for any given horizon length T. Further discussion can be found in Secomandi (2008), which shows that for a general class of network revenue management problems, re-solving does not guarantee improvement (even when the exact demand model is known). In particular, Secomandi (2008) finds that lack of sequential consistency (i.e. when a previous solution is no longer feasible upon re-solving) may lead to poor re-solving behavior.

3.2. Numerical Results

In this section, we first numerically analyze the performance of the TS-fixed and TS-update algorithms for the setting where a single product is sold throughout the course of the selling season, and we compare these results to other proposed algorithms in the literature. Then we present a numerical analysis for a multi-product example; for consistency, the example we chose to use is identical to the one presented in Section 3.4 of Besbes and Zeevi (2012).

3.2.1. Single Product Example

Consider a retailer who sells a single product ($N = 1$) throughout a finite selling season. Without loss of generality, we can assume that the product is itself the resource ($M = 1$) which has limited inventory. The set of feasible prices is $\{\$29.90, \$34.90, \$39.90, \$44.90\}$, and the mean demand is given by $d(\$29.90) = 0.8, d(\$34.90) = 0.6, d(\$39.90) = 0.3$, and $d(\$44.90) = 0.1$. As common in revenue management literature, we show numerical results in an asymptotic regime when inventory is scaled linearly with time: initial inventory $I = \alpha T$, for $\alpha = 0.25$ and 0.5.

We evaluate and compare the performance of the following five dynamic pricing algorithms which have been proposed for our setting:

- TS-fixed: defined in Algorithm 1. We use the independent Beta prior as in Example 1.
- TS-update: defined in Algorithm 2. We use the independent Beta prior as in Example 1.
- BZ: the algorithm proposed in Besbes and Zeevi (2012), which first explores all prices and then exploits the best pricing strategy by solving a linear program once. In our implementation, we divide the exploration and exploitation phases at period $\tau = T^{2/3}$, as suggested in their paper.
- PD-BwK: the algorithm proposed in Badanidiyuru et al. (2013) that is based on a primal-dual algorithm to solve LP($d(t)$) and uses the UCB algorithm to estimate demand. For each period, it estimates upper bounds on revenue, lower bounds on resource consumption, and the dual price of each resource, and then selects the price vector with the highest revenue-to-resource-price ratio.
- TS: this is the original Thompson sampling algorithm described in Thompson (1933), which has been proposed for use as a dynamic pricing algorithm but does *not* consider inventory constraints; see Appendix.

We measure performance as the average percent of "optimal revenue" achieved over 500 simulations. By "optimal revenue", we are referring to the upper bound on optimal revenue where the retailer knows the mean demand at each price prior to the selling season; this upper bound is the optimal value of LP(d), described in Section 3.1.1. Thus, the percent of the true optimal revenue achieved is at least as high as the numbers shown. Figure 1 shows performance results for the five algorithms outlined above.

The first thing to notice is that all four algorithms that incorporate inventory constraints converge to the optimal revenue as the length of the selling season increases. The TS algorithm, which does not incorporate inventory constraints, does not converge to the optimal revenue. This is because in each of the examples shown, the optimal pricing strategy of LP(d) is a mixed strategy where two prices are offered throughout the selling season as opposed to a single price being offered to all customers. The optimal strategy of LP(d) when $I = 0.25T$ is to offer the product at \$39.90 to

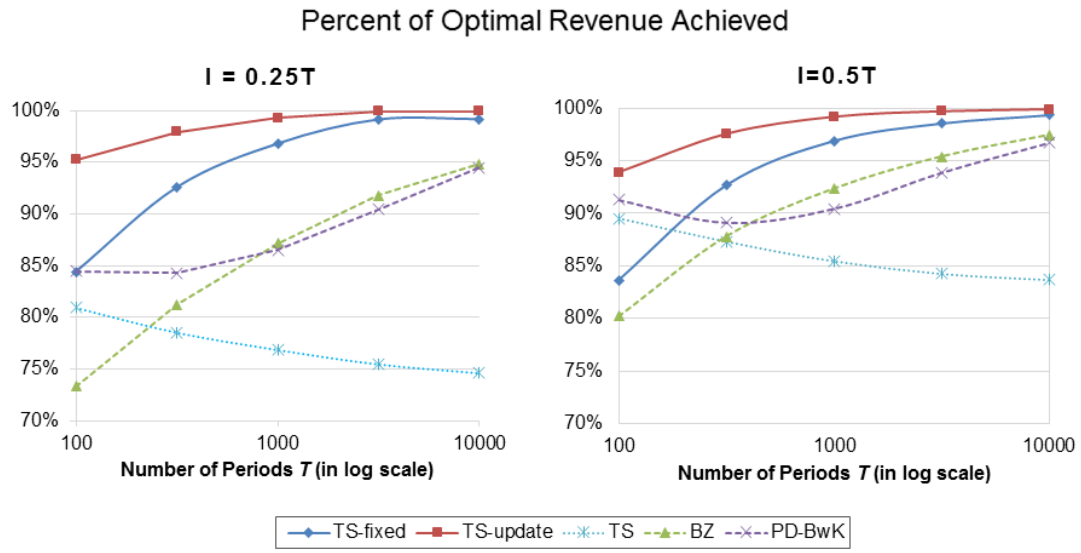

Figure 1 Performance Comparison of Dynamic Pricing Algorithms: Single Product Example

$\frac{3}{4}$ of the customers and \$44.90 to the remaining $\frac{1}{4}$ of the customers. The optimal strategy when $I = 0.5T$ is to offer the product at \$34.90 to $\frac{2}{3}$ of the customers and \$39.90 to the remaining $\frac{1}{3}$ of the customers. In both cases, TS converges to the suboptimal price \$29.90 offered to all the customers since this is the price that maximizes expected revenue given unlimited inventory. This really highlights the necessity of incorporating inventory constraints when developing dynamic pricing algorithms. More generally, this highlights the necessity of incorporating operational constraints when adapting machine learning algorithms for operational use.

Second, we note that in this example, TS-update outperforms all of the other algorithms in every scenario, while TS-fixed ranks second in most cases. Interestingly, when considering only those algorithms that incorporate inventory constraints, the gap between TS-update and the others generally increases when (i) the length of the selling season is short, and (ii) the ratio I/T is small. This is consistent with many other examples that we have tested and suggests that TS-update is particularly powerful (as compared to the other algorithms) when inventory is very limited and the selling season is short. In other words, TS-update is able to more quickly learn mean demand and identify the optimal pricing strategy, which is particularly useful for low inventory settings.

3.2.2. Multi-Product Example

Now we consider an example used by Besbes and Zeevi (2012) where a retailer sells two products ($N = 2$) using three resources ($M = 3$). Selling one unit of product $i = 1$ consumes 1 unit of resource $j = 1$, 3 units of resource $j = 2$, and no units of resource $j = 3$. Selling one unit of product $i = 2$ consumes 1 unit of resource 1, 1 unit of resource 2, and 5 units of resource 3. The set of feasible

prices is $(p_1, p_2) \in \{(1, 1.5), (1, 2), (2, 3), (4, 4), (4, 6.5)\}$. Besbes and Zeevi (2012) assume customers arrive according to a multivariate Poisson process. They considered the following three possibilities for mean demand of each product as a function of the price vector:

1. *Linear:* $\mu(p_1, p_2) = (8 - 1.5p_1, 9 - 3p_2)$,
2. *Exponential:* $\mu(p_1, p_2) = (5e^{-0.5p_1}, 9e^{-p_2})$, and
3. *Logit:* $\mu(p_1, p_2) = \left(\frac{10e^{-p_1}}{1+e^{-p_1}+e^{-p_2}}, \frac{10e^{-p_2}}{1+e^{-p_1}+e^{-p_2}}\right)$.

We compare BZ, TS-fixed and TS-update for this example. We use the independent Gamma prior described in Example 2. Since the PD-BwK algorithm proposed in Badanidiyuru et al. (2013) does not apply to the setting where customers arrive according to a Poisson process, we did not include this algorithm in our comparison.

We again measure performance as the average percent of "optimal revenue" achieved, where optimal revenue refers to the upper bound on optimal revenue when the retailer knows the mean demand at each price prior to the selling season. Thus, the percent of optimal revenue achieved is at least as high as the numbers shown. Figure 2 shows average performance results over 500 simulations for each of the three underlying demand functions; we show results when inventory is scaled linearly with time, *i.e.* initial inventory $I = \alpha T$, for $\alpha = (3, 5, 7)$ and $\alpha = (15, 12, 30)$.

As in the single product example, each algorithm converges to the optimal revenue as the length of the selling season increases. In most cases, TS-update and TS-fixed outperform the BZ algorithm proposed in Besbes and Zeevi (2012). TS-fixed has slightly worse performance than TS-update in most cases, but in a few cases the difference between the two algorithms is almost indistinguishable. For each set of parameters and when $T = 10,000$, TS-update and TS-fixed achieve 99–100% of the optimal revenue whereas the Besbes and Zeevi (2012) algorithm achieves 92–98% of the optimal revenue. As we saw in the single product example, TS-update performs particularly well when inventory is very limited ($I = (3, 5, 7)T$); it is able to more quickly learn mean demand and identify the optimal pricing strategy. TS-update and TS-fixed also seem to perform particularly well when mean demand is linear. Finally, note that the algorithm's performance appears to be fairly consistent across the three demand models tested.

4. Extensions and Further Applications

Thus far, we have developed simple and effective algorithms that dynamically change prices to learn demand and maximize revenue, which can be applied to a practical setting faced by many retailers. The algorithms build directly from the popular machine learning algorithm known as Thompson sampling by inserting an additional linear programming subroutine to incorporate inventory constraints. In this section, we show the broader applicability of this approach, highlighting

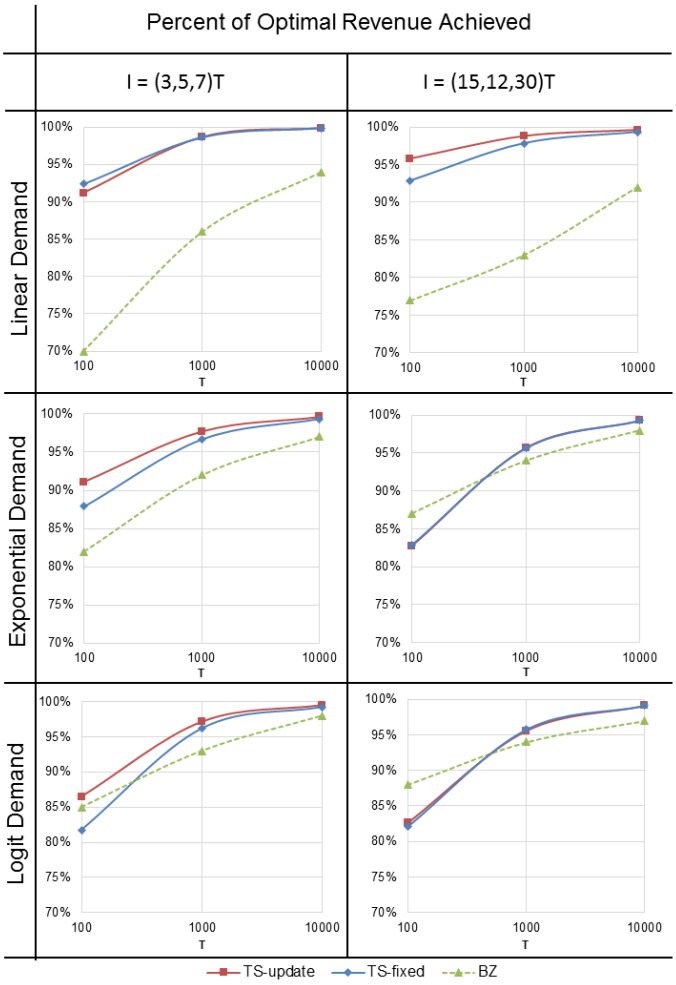

Figure 2 Performance Comparison of Dynamic Pricing Algorithms – Multi-Product Example

how Thompson sampling can be adapted to address a variety of additional revenue management and operations management challenges.

For brevity, we only present algorithms that do not update inventory information, analogous to TS-fixed. One can easily modify the algorithms below to include inventory updating analogous to TS-update by replacing constant c_j with $c_j(t)$.

4.1. Continuous Price Sets with Linear Demand

The algorithms proposed in Section 2 were designed specifically for the setting with a discrete set of prices. In this section, we show how our algorithms can be adapted to a model setting where the retailer can choose prices from a continuous price set.

Let $P(t) = [P_i(t)]_{i\in[N]}$ be the price vector that the retailer offers in period t. We require $P(t) \in \mathcal{P}$, where $\mathcal{P}$ is a bounded polyhedral set (*i.e.*, a polytope) representing all feasible prices. We consider

linear demand functions as in Keskin and Zeevi (2014), and extend their model setting to include inventory constraints. Let $D(t) = [D_i(t)]_{i\in N}$ be the demand in period t. We assume that there is a vector $\alpha \in \mathbb{R}^N$ and a parameter matrix $B \in \mathbb{R}^{N\times N}$, such that $D(t) = \alpha + BP(t) + \epsilon(t)$, where $\epsilon(t) \in \mathbb{R}^N$ is demand noise. We assume $\epsilon(t)$ is sampled independently from a known multivariate normal distribution with zero mean. The demand parameter $\theta = (\alpha, B) \in \Theta \subset \mathbb{R}^{N\times(N+1)}$ is unknown, while the retailer has a prior distribution of θ over Θ. We also assume that Θ is bounded; for any $\theta' = (\alpha', B') \in \Theta$, $-B'$ is positive definite, and there exists $p \in \mathcal{P}$ such that $\alpha' + B'p \leq 0$, where "$\leq$" holds element-wise.

As described in Section 1.1, a key distinction between discrete vs. continuous price sets arises from the structure of their respective revenue optimization problems. If the price set is discrete and the retailer faces inventory constraints, there may not exist any single price that maximizes revenue asymptotically as the number of periods T increases; the retailer must maximize over *distributions of prices* as in TS-fixed and TS-update. In contrast, if the price set is continuous and the demand function satisfies certain regularity conditions (Gallego and Van Ryzin 1994), there always exists a *single* price that is asymptotically optimal, regardless of the presence of inventory constraints.

We next present TS-linear (Algorithm 3) – a natural adaptation of our Thompson sampling based algorithms in Section 2 to this new model setting.

Algorithm 3 Thompson Sampling with Inventory for Linear Demand (TS-linear)

Repeat the following steps for all periods $t \in [T]$:

1. *Sample Demand*: Sample a random parameter vector $\theta(t) = (\alpha(t), B(t)) \in \Theta$ from the posterior distribution of θ given history $\mathcal{H}_{t-1}$.

2. *Optimize Prices given Sampled Demand*: Solve the following quadratic programing problem:

$$\begin{aligned} \mathsf{QP}(\theta(t)): \quad \max_{p} \quad & p^T\left(\alpha(t) + B(t)p\right) \\ \text{subject to} \quad & A^T(\alpha(t) + B(t)p) \leq c \\ & p \in \mathcal{P}. \end{aligned}$$

3. *Offer Price*: Let $P(t)$ be the optimal solution to $\mathsf{QP}(\theta(t))$; offer price vector $P(t)$.

4. *Update Estimate of Parameter*: Observe demand $D(t)$. Update the history $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{P(t), D(t)\}$ and the posterior distribution of θ given $\mathcal{H}_t$.

Step 1 of TS-linear is similar to step 1 in TS-fixed. In problem $\mathsf{QP}(\theta(t))$ of step 2, recall that $c = (I_j/T)_{j\in[M]}$ is a vector representing the average inventory per period over the entire selling horizon. Matrix $A = [a_{ij}]_{i\in[N],j\in[M]}$ is the resource consumption matrix. Note that unlike $\mathsf{LP}(d(t))$

in TS-fixed, the decision variables of $\mathsf{QP}(\theta(t))$ are single prices rather than probability distributions over prices, for the reason that we discussed earlier. By our model assumptions, $\mathsf{QP}(\theta(t))$ is always feasible and has a concave objective function, so the optimization step is well defined and can be performed efficiently. Finally, steps 3 and 4 are similar to those in TS-fixed.

Performance Analysis For TS-linear, we have the following performance guarantee.

THEOREM 3. *The Bayesian regret of TS-linear is bounded by*

$$\text{BayesRegret}(T) \leq O(N^2 \log T \sqrt{T}).$$

Unlike Theorems 1 and 2, the regret bound of TS-linear does not depend on the number of price vectors (which is uncountable in this case). Instead, the regret bound in Theorem 3 depends on the number of unknown parameters, which is equal to $N^2 + N = O(N^2)$. The regret bound is near optimal up to a $\log T$ factor, since for a linear bandit problem with $\Omega(N^2)$ parameters – a special case of our linear demand model without inventory constraints – no algorithm can achieve a Bayesian regret bound smaller than $\Omega(N^2\sqrt{T})$ (see Theorem 2.1, Rusmevichientong and Tsitsiklis (2010)).

The proof outline of Theorem 3 follows the proof idea in Theorem 1 (TS-fixed). In the proof for Theorem 1, we build a connection between the dynamic pricing problem and the multi-armed bandit problem – a special case of the finite price model without inventory constraints. Similarly, in the proof for Theorem 3, we build a connection between our problem and the *linear bandit problem* (Rusmevichientong and Tsitsiklis 2010, Russo and Van Roy 2014), which as we mentioned earlier is a special case of our linear demand model without inventory constraints. The detailed proof of Theorem 3 can be found in the E-companion.

4.2. Contextual Network Revenue Management

In the model of Section 2.1, we assumed that demand is i.i.d. given any price vector. In practice, the retailer often knows some exogenous information that may affect demand, *e.g.* including seasonality, changes of competitors' prices, and customer attributes. In this section, we extend our model and algorithms to a setting that can incorporate such contextual information.

Model and Algorithm We extend the model setting in Section 2.1 based on the contextual bandits model in Badanidiyuru et al. (2014). Suppose that at the beginning of each period $t \in [T]$, the retailer observes some *context* (or *feature*) $\xi(t)$, where $\xi(t)$ belongs to some discrete set $\mathcal{X}$ (*i.e.*, the feature space). We assume that context $\xi(t)$ is sampled i.i.d. from a *known* distribution. The retailer then observes $\xi(t)$ and selects a price vector $P(t) \in \{p_1, p_2, \ldots, p_K, p_\infty\}$, where p_∞ forces demand to be zero with probability one. We assume that for any $\xi \in \mathcal{X}$, the product demand under

a given price vector is i.i.d., parameterized by an *unknown* vector $\theta \in \Theta$. In particular, we denote by $d_{ik}(\xi \mid \theta)$ the mean demand of product $i \in [N]$ under price vector p_k ($\forall k \in [K]$), given context ξ and parameter θ. The retailer is given a prior distribution of θ at the beginning of the selling season, and maximizes expected revenue over the course of the selling season.

Our assumption that the retailer knows the distribution of context is realistic in many practical applications. For example, suppose Amazon wants to offer different prices to its Prime members and non-members. We can then define context $\xi(t)$ as a dummy variable indicating whether the t^{th} arriving customer is a Prime member. Amazon likely has a pretty good estimate of the percentage of customers visiting its website who are Prime members based on historical data; thus we assume they know this distribution of prime members in our model.

Algorithm 4 Thompson Sampling with Inventory and Contextual Information (TS-contextual)

Repeat the following steps for all periods $t = 1, ..., T$:

1. *Sample Demand*: Sample $\theta(t)$ from the posterior distribution of θ given history $\mathcal{H}_{t-1}$.

2. *Optimize Prices given Sampled Demand*: Solve the following linear program, denoted by $\mathsf{LP}(\xi \mid \theta(t))$:

$$\begin{aligned} \mathsf{LP}(\xi \mid \theta(t)): \quad \max_x \mathrm{E}_\xi \left[\sum_{k=1}^{K} \sum_{i=1}^{N} p_{ik} d_{ik}(\xi \mid \theta(t)) x_{\xi,k} \right] \\ \text{subject to } \mathrm{E}_\xi \left[\sum_{k=1}^{K} \sum_{i=1}^{N} a_{ij} d_{ik}(\xi \mid \theta(t)) x_{\xi,k} \right] \le c_j, \ \forall j \in [M] \\ \sum_{k=1}^{K} x_{\xi,k} \le 1, \ \forall \xi \in \mathcal{X} \\ x_{\xi,k} \ge 0, \ \forall \xi \in \mathcal{X}, k \in [K]. \end{aligned}$$

Let $x(t) = (x_{\xi,k}(t))_{\xi \in \mathcal{X}, k \in [K]}$ be the optimal solution to $\mathsf{LP}(\xi \mid \theta(t))$.

3. *Offer Price*: Observe the realized context $\xi(t)$, and then offer price vector $P(t) = p_k$ with probability $x_{\xi(t),k}(t)$, and choose $P(t) = p_\infty$ with probability $1 - \sum_{k=1}^{K} x_{\xi(t),k}(t)$.

4. *Update Estimate of Parameter*: Observe demand $D(t)$. Update the history $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{\xi(t), P(t), D(t)\}$ and the posterior distribution of θ.

For this model, we propose the TS-contextual algorithm which is an extension of TS-fixed that incorporates the contextual information; TS-contextual is presented in Algorithm 4. In step 1, the retailer samples an instance of the demand parameter from its posterior distribution, similar to step 1 in TS-fixed. The linear program in step 2 has to be modified to include the context information. We use $\mathrm{E}_\xi[\cdot]$ to denote the expectation operator where ξ is a random variable following

the probability distribution of contexts. Thus, the objective and the constraints of the LP account for the randomness of contexts appearing in future periods. The decision variable x is a family of distributions over $k \in [K]$ for each context $\xi \in \mathcal{X}$. In step 3, the retailer observes the context and then chooses price according to the probability distribution associated with this particular context. Finally, step 4 is similar to that in TS-fixed.

Performance Analysis The following theorem provides a Bayesian regret bound for TS-contextual.

THEOREM 4. *The Bayesian regret of TS-contextual is bounded by*

$$\text{BayesRegret}(T) \leq \left(18p_{\max} + 38\sum_{j=1}^{M}\sum_{i=1}^{N} p_{\max}^{j} a_{ij}\bar{d}_i\right) \cdot \sqrt{|\mathcal{X}|TK\log K},$$

where constants $p_{\max}$ and $p_{\max}^{j}$ are defined in Section 3.1.2.

The proof outline for Theorem 4 closely follows the proof for Theorem 1. In fact, Theorem 1 can be viewed as a special case of Theorem 4 when $|\mathcal{X}| = 1$. The detailed proof of Theorem 4 can be found in the E-companion.

REMARK 4. The regret bound depends on the size of the feature space by $O(\sqrt{|\mathcal{X}|})$, and is meaningful when $|\mathcal{X}|$ is small compared to T. Recently, for a similar problem, Agrawal et al. (2016) shows that one can establish a regret bound of $O(\sqrt{KT\log(T|\Pi|)})$, where Π is the set of admissable policies that maps $\mathcal{X}$ to $\{p_1, \ldots, p_K\}$. If Π includes all policies, then $|\Pi| = K^{|\mathcal{X}|}$, and we recover the regret bound in Theorem 4 (up to log factors).

4.3. Bandits with Knapsacks Problem

Beyond pricing, our algorithms can be applied to other operations management challenges. Badanidiyuru et al. (2013) propose a "bandits with knapsacks" problem, which can model various applications in pricing, crowdsourcing and advertising.

The "bandits with knapsacks" problem is the following: there are K arms representing different "actions" that the decision maker can take, and M resources with limited budget. At each time period $t \in [T]$, the decision maker chooses to pull one of the arms. If an arm is pulled, the decision maker receives a random reward in $[0, 1]$, and consumes a certain amount of resources, represented by a random vector in $[0, 1]^M$. For any fixed arm, the reward and resource consumption vector is sampled i.i.d. from a fixed but unknown distribution on $[0, 1] \times [0, 1]^M$. The process stops at the first time when the total consumption of some resource exceeds its budget; otherwise, the process stops at period T.

Next we give an example for applying the "bandits with knapsacks" model to display advertising.

Example Application: Display Advertising Consider an online ad platform (*e.g.* Google). For each user logging on to a third-party website (*e.g.* The New York Times), Google may display a banner ad on that website. If the user clicks the ad, the advertiser sponsoring the ad pays a fixed amount of money that is split between Google and the publisher (*i.e.*, the New York Times). If the user does not click the ad, no payment is made. Google then faces a problem of allocating ads to publishers to maximize its revenue, while satisfying advertisers' budgets. Assuming that click rates for ads are unknown, this example fits into the bandits with knapsacks model where each arm is an ad, and each resource corresponds to an advertiser's budget. If an ad is clicked, Google receives a reward and consumes some of the advertiser's budget.

We propose TS-BwK, a Thompson sampling based algorithm for the Bandits with Knapsacks problem. Let r_k be the mean reward of arm k, and let b_{jk} be the expected consumption of resource j when arm k is pulled. (For the advertising example above, r_k is the mean revenue when ad k is displayed, and b_{jk} is the expected payment from advertiser j when ad k is displayed.) We denote by $A(t)$ the arm that is pulled at time t, and denote by $R(t)$ the observed reward and $B(t) \in [0,1]^M$ the observed resource consumption vector at time t. For each resource $j \in [M]$, we use $c_j := I_j/T$ to denote the average units of resource per time period. Algorithm 5 outlines the steps of TS-BwK.

In step 1, for each arm, the firm samples mean reward and expected consumption of each resource; this is analogous to the retailer sampling mean demand (equivalently, revenue) in step 1 of TS-fixed. In step 2, the firm maximizes its reward subject to resource constraints, just as the retailer maximizes revenue subject to inventory constraints in step 2 of TS-fixed. In step 3, the firm chooses the arm to maximize its reward; this is analogous to the retailer offering the price to maximize its revenue in step 3 of TS-fixed. Finally in step 4, the firm observes the reward and updates history, just as the retailer observes demand and updates history in step 4 of TS-fixed.

Performance Analysis The following theorem shows that TS-BwK has a Bayesian regret bounded by $O(\sqrt{T}\log T)$.

THEOREM 5. *The Bayesian regret of TS-BwK is bounded by*

$$\text{BayesRegret}(T) \leq \left(\frac{37M}{c_{\min}}\log(T) + 20\right)\sqrt{KT\log K},$$

where $c_{\min} = \min_{j\in[M]} c_j$.

The proof of Theorem 5 can be found in the E-companion.

REMARK 5. Compared to the $O(\sqrt{KT\log K})$ regret bound for TS-fixed, the regret bound for TS-BwK has an extra $\log(T)$ factor. This is due to a distinction in model assumptions: In the network

Algorithm 5 Thompson Sampling for Bandits with Knapsacks (TS-BwK)

Repeat the following steps for all periods $t = 1, ..., T$:

1. *Sample*: Sample a vector $(r_k(t), b_{jk}(t))_{k\in[K],j\in[M]}$ from the posterior distribution of $(r_k, b_{jk})_{k\in[K],j\in[M]}$ given observed history $\mathcal{H}_{t-1}$.

2. *Optimize*: Solve the following linear program:

$$\begin{aligned} \max_x \ & \sum_{k=1}^{K} r_k(t) x_k \\ \text{subject to } & \sum_{k=1}^{K} b_{jk}(t) x_k \le c_j, \ \forall j \in [M] \\ & \sum_{k=1}^{K} x_k \le 1 \\ & x_k \ge 0, \ k \in [K]. \end{aligned}$$

Let $x(t) = (x_1(t), \ldots, x_K(t))$ be its optimal solution.

3. *Take Action*: Pull arm k with probability $x_k(t)$, and take no action with probability $1 - \sum_{k=1}^{K} x_k(t)$.

4. *Update*: If an arm is pulled, observe reward $R(t)$ and consumption vector $B(t)$. Update history $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{A(t), R(t), B(t)\}$.

revenue management model, if some resource is depleted, we allow the selling process to continue and the retailer can still offer products that do not consume this particular resource. In the bandits with knapsacks problem defined by Badanidiyuru et al. (2013), the process immediately stops when the total consumption of *any* resource exceeds its budget. Due to this distinction, additional proof steps are required for TS-BwK, which result in an extra $\log(T)$ factor.

5. Conclusion

We focus on a finite-horizon, price-based network revenue management problem in which an online retailer aims to maximize revenue from multiple products with limited inventory. As common in practice, the retailer does not know the mean demand for each price, but can learn mean demand by observing customer purchase decisions over time. The main contribution of our work is the development of implementable, effective dynamic pricing algorithms which balance the exploration-exploitation tradeoff to learn mean demand and maximize revenue over the course of the selling season in the presence of limited inventory constraints. These algorithms show the power of extending a machine learning technique, Thompson sampling, for practical use in revenue management by incorporating inventory constraints into the retailer's pricing strategy using a linear programming subroutine. Analogous to the classical Thompson sampling algorithm, in every time period, our

algorithms update the posterior distribution of the unknown demand model parameter and then use a randomly sampled parameter value from the posterior distribution to make pricing decisions.

We show that our $O(\sqrt{T})$ Bayesian regret bound matches the best possible lower bound $\Omega(\sqrt{T})$. Furthermore, our algorithms prove to have promising numerical performance results when compared to other algorithms developed for the same setting. We also show how the algorithms can be extended to model settings with continuous price sets or contextual information. More broadly, our algorithms can be adapted to a general multi-armed bandit problem with resource constraints, with applications to other operations management challenges beyond pricing, such as display advertising.

There are several directions for future work that we think would be valuable. A first direction of future work is to extend our algorithms for assortment optimization, another class of important revenue management problems. Second, one future direction is generalizing our TS-contextual algorithm to avoid the regret's dependence on the size of the feature space. A third direction involves developing and conducting field experiments in partnership with a retailer(s) to understand the effectiveness of our algorithms in practice.

Endnotes

1. Although Chen et al. (2014) also uses inventory update in a price-based revenue management problem with unknown demand, their algorithm separates the selling season into exploration and exploitation phases, and applies inventory update only in the exploitation phase; in contrast, TS-update applies inventory updating throughout the entire selling horizon.

2. The boundedness of demand is required since our proof applies a large deviation bound and analysis technique by Audibert and Bubeck (2009) and Bubeck and Liu (2013) that is specific to bounded probability distributions. In general, if demand is unbounded and follows a subexponential distribution, one can replace the large deviation bound in the proof by subexponential tail bounds, and replace the demand upper bound $\bar{d}_i$ by an upper bound on the mean demand. The same proof in the appendix works and would lead to $O(\sqrt{TK\log T})$ regret bounds in both Theorems 1 and 2.

Acknowledgments

The authors thank the Area Editor, the Associate Editor and three anonymous referees for comments that have greatly improved the paper. This work was supported in part by Accenture through the Accenture-MIT Alliance in Business Analytics.

Appendix. Multi-Armed Bandit Problem and Thompson Sampling with Unlimited Inventory

The original Thompson sampling algorithm was proposed by Thompson (1933) for the multi-armed bandit problem without inventory constraints. The problem is the following: Suppose there are K arms. At each time period $t \in [T]$, the decision maker chooses to pull one of the arms, and receives a random reward $R(t) \in [0,1]$. We assume that if arm $k \in [K]$ is pulled, the reward is sampled i.i.d. from a fixed distribution with mean r_k. The true value of $r = (r_k)_{k\in[K]}$ is unknown, while the decision maker only knows a prior distribution of r on $[0,1]^K$. For this problem, Thompson (1933) proposed the following algorithm (Algorithm 6).

Algorithm 6 Thompson Sampling with Unlimited Inventory

Repeat the following steps for each period $t = 1, ..., T$:

1. *Sample*: Sample a vector $(\theta_k(t))_{k\in[K]}$ from the posterior distribution of r given $\mathcal{H}_{t-1}$.
2. *Take Action*: Select an arm $A(t) \in \arg\max_{k\in[K]} \theta_k(t)$.
3. *Update*: Observe reward $R(t)$ and update history $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{A(t), R(t)\}$.

References

Agrawal, S., N. R. Devanur, L. Li, N. Rangarajan. 2016. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. *29th Conf. on Learning Theory (COLT)*. 4–18.

Agrawal, S., N. Goyal. 2013. Further optimal regret bounds for thompson sampling. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. 99–107.

Araman, V. F., R. Caldentey. 2009. Dynamic pricing for nonperishable products with demand learning. *Operations Research* **57**(5) 1169–1188.

Audibert, J.-Y., S. Bubeck. 2009. Minimax policies for adversarial and stochastic bandits. *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*. 217–226.

Auer, P., N. Cesa-Bianchi, P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47**(2-3) 235–256.

Aviv, Y., G. Vulcano. 2012. Dynamic list pricing. Özalp Özer, Robert Phillips, eds., *The Oxford Handbook of Pricing Management*. Oxford University Press, Oxford, 522–584.

Badanidiyuru, A., R. Kleinberg, A. Slivkins. 2013. Bandits with knapsacks. *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*. 207–216.

Badanidiyuru, A., J. Langford, A. Slivkins. 2014. Resourceful contextual bandits. *26th Conf. on Learning Theory (COLT)*. 1109–1134.

Besbes, O., A. Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* **57**(6) 1407–1420.

Besbes, O., A. Zeevi. 2012. Blind network revenue management. *Operations Research* **60**(6) 1537–1550.

Bubeck, S., N. Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* **5**(1) 1–122.

Bubeck, S., C.-Y. Liu. 2013. Prior-free and prior-dependent regret bounds for thompson sampling. *Advances in Neural Information Processing Systems*. 638–646.

Chapelle, O., L. Li. 2011. An empirical evaluation of thompson sampling. *Advances in Neural Information Processing Systems*. 2249–2257.

Chen, Q., S. Jasin, I. Duenyas. 2014. Adaptive parametric and nonparametric multi-product pricing via self-adjusting controls. Working paper, Ross School of Business, University of Michigan.

Chen, Y., V. F. Farias. 2013. Simple policies for dynamic pricing with imperfect forecasts. *Operations Research* **61**(3) 612–624.

Cooper, W. L. 2002. Asymptotic behavior of an allocation policy for revenue management. *Operations Research* **50**(4) 720–727.

den Boer, A. V. 2015. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science* **20**(1) 1–18.

Farias, V., B. Van Roy. 2010. Dynamic pricing with a prior on market response. *Operations Research* **58**(1) 16–29.

Freedman, D. A. 1963. On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics* **34**(4) 1386–1403.

Gallego, G., G. Van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science* **40**(8) 999–1020.

Gallego, G., G. Van Ryzin. 1997. A multiproduct dynamic pricing problem and its applications to network yield management. *Operations Research* **45**(1) 24–41.

Harrison, J., N. Keskin, A. Zeevi. 2012. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science* **58**(3) 570–586.

Jasin, S. 2015. Performance of an lp-based control for revenue management with unknown demand parameters. *Operations Research* **63**(4) 909–915.

Jasin, S., S. Kumar. 2012. A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research* **37**(2) 313–345.

Kaufmann, E., N. Korda, R. Munos. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. *Algorithmic Learning Theory*. Springer, 199–213.

Keskin, N. B., A. Zeevi. 2014. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research* **62**(5) 1142–1167.

Lerman, S. 2014. E-commerce and online auctions in the us. Tech. rep., IBISWorld Industry Report 45411a.

Özer, Ö., R. Phillips. 2012. *The Oxford Handbook of Pricing Management*. Oxford University Press.

Rusmevichientong, P., J. N. Tsitsiklis. 2010. Linearly parameterized bandits. *Mathematics of Operations Research* **35**(2) 395–411.

Russo, D., B. Van Roy. 2014. Learning to optimize via posterior sampling. *Mathematics of Operations Research* **39**(4) 1221–1243.

Secomandi, N. 2008. An analysis of the control-algorithm re-solving issue in inventory and revenue management. *Manufacturing & Service Operations Management* **10**(3) 468–483.

Talluri, K. T., G. J. van Ryzin. 2005. *Theory and Practice of Revenue Management*. Springer-Verlag.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**(3/4) 285–294.

Wang, Z., S. Deng, Y. Ye. 2014. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research* **62**(2) 318–331.

E-companion: Theoretical Analysis for "Online Network Revenue Management using Thompson Sampling"

EC.1. Proof for TS-fixed

THEOREM 1. *The Bayesian regret of* TS-fixed *is bounded by*

$$\text{BayesRegret}(T) \le \left(18p_{\max} + 37\sum_{i=1}^{N}\sum_{j=1}^{M} p^j_{\max} a_{ij}\bar{d}_i\right)\cdot\sqrt{TK\log K}.$$

Proof. Let $d = \{d_{ik}\}_{i\in[N],k\in[K]}$ be the mean demand under the true parameter θ. For each price $k\in[K]$, we denote by $r_k = \sum_{i=1}^N p_{ik}d_{ik}$ the mean revenue, and denote by $b_{jk} = \sum_{i=1}^N a_{ij}d_{ik}$ the expected consumption of resource j. Recall that x^* is the optimal solution of the following LP:

$$\begin{aligned}\mathsf{LP}(d):\quad \max_x\ & \sum_{k=1}^K r_k x_k\\ \text{subject to}\ & \sum_{k=1}^K b_{jk}x_k \le c_j, \quad \forall j\in[M]\\ & \sum_{k=1}^K x_k \le 1\\ & x_k \ge 0, \quad \forall k\in[K].\end{aligned}$$

Similarly, let the mean demand under $\theta(t)$, the sampled parameter at period t, be $d(t) = \{d_{ik}(t)\}_{i\in[N],k\in[K]}$. We denote by $r_k(t) = \sum_{i=1}^N p_{ik}d_{ik}(t)$ and $b_{jk}(t) = \sum_{i=1}^N a_{ij}d_{ik}(t)$ the revenue and resource consumption under the demand sampled at period t. Recall that $x(t)$ is the optimal solution of the following LP:

$$\begin{aligned}\mathsf{LP}(d(t)):\quad \max_x\ & \sum_{k=1}^K r_k(t) x_k\\ \text{subject to}\ & \sum_{k=1}^K b_{jk}(t)x_k \le c_j, \quad \forall j\in[M]\\ & \sum_{k=1}^K x_k \le 1\\ & x_k \ge 0, \quad \forall k\in[K].\end{aligned}$$

If we assume unlimited inventory, the expected revenue of TS-fixed over the selling horizon is given by

$$\mathrm{E}[\sum_{t=1}^T\sum_{k=1}^K r_k \mathbf{1}_{\{P(t)=p_k\}}] = \mathrm{E}[\sum_{t=1}^T\sum_{k=1}^K r_k \mathrm{E}[\mathbf{1}_{\{P(t)=p_k\}}\mid\theta,\theta(t)]] = \mathrm{E}[\sum_{t=1}^T\sum_{k=1}^K r_k x_k(t)]. \tag{EC.1}$$

Of course, to calculate the actual revenue, we should subtract from Eq (EC.1) the amount of revenue associated with lost sales. We claim that this amount is no more than

$$\sum_{j=1}^M p^j_{\max}\mathrm{E}\left[\left(\sum_{i=1}^N\sum_{t=1}^T a_{ij}D_i(t) - I_j\right)^+\right], \tag{EC.2}$$

where $(x)^+ = \max\{x,0\}$ and $p^j_{\max} = \max_{i:a_{ij}\neq 0,k\in[K]}(p_{ik}/a_{ij})$. To see this, recall that $D_i(t)$ is the realized demand of product i at time t, so $\left(\sum_{i=1}^N\sum_{t=1}^T a_{ij}D_i(t) - I_j\right)^+$ is the amount of resource j consumed beyond

the inventory budget I_j. The coefficient $\max_{i:a_{ij}\neq 0,k\in[K]}(p_{ik}/a_{ij})$ is the maximum revenue that can be achieved by adding one unit of resource j. Therefore, Eq (EC.2) is an upper bound of the revenue that should be subtracted from (EC.1).

In Section 3.1.1, we have shown that

$$\text{BayesRegret}(T) \leq T \cdot \text{E}[\sum_{k=1}^{K} r_k x_k^*] - \text{E}[\text{Rev}(T)].$$

Therefore, by Eq (EC.1) and (EC.2), we have

$$\begin{aligned}
&\text{BayesRegret}(T)\\
\leq& T \cdot \text{E}[\sum_{k=1}^{K} r_k x_k^*] - \text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k x_k(t)] + \sum_{j=1}^{M} p_{\max}^j \text{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} a_{ij} D_i(t) - I_j\right)^+\right]\\
=& \underbrace{\text{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k (x_k^* - x_k(t))\right]}_{\text{(I)}} + \underbrace{\sum_{j=1}^{M} p_{\max}^j \text{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} a_{ij} D_i(t) - I_j\right)^+\right]}_{\text{(II)}}.
\end{aligned}$$

To complete the proof, we show that both (I) and (II) are bounded by $O(\sqrt{T})$.

Part (I): Revenue Bound. We first show that

$$\text{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k (x_k^* - x_k(t))\right] \leq p_{\max} \cdot 18\sqrt{KT\log(K)}, \tag{EC.3}$$

where $p_{\max} = \max_{k\in[K]}\{\sum_{i=1}^{N} p_{ik}\bar{d}_i\}$ is the maximum revenue that can be achieved in a single period.

Let $\mathcal{H}_{t-1} = (P(1), D(1), \ldots, P(t-1), D(t-1))$ be the history available at the beginning of period t. In the TS-fixed algorithm, note that the values of $\theta(t)$ are sampled from the posterior distribution of θ conditional on history $\mathcal{H}_{t-1}$, *i.e.* $\mathbb{P}(\theta \mid \mathcal{H}_{t-1}) = \mathbb{P}(\theta(t) \mid \mathcal{H}_{t-1})$. In order to see this equality more clearly, note that since Thompson sampling is defined in a Bayesian setting, the value of the unknown parameter (or parameters) θ is a random variable. Nature draws θ at the beginning of the horizon, but the value is not revealed to the retailer, so the retailer keeps updating its estimation using a posterior distribution $P(\theta|H_{t-1})$. Then, in each period, the retailer draws a random parameter $\theta(t)$ according to the posterior distribution. So, although θ does not necessarily equal $\theta(t)$, by definition, the two random variables have the same *probability distribution* and thus $P(\theta|H_{t-1}) = P(\theta(t)|H_{t-1})$. This important equality is also used in Russo and Van Roy (2014).

We denote by d the mean demand under θ and $d(t)$ the mean demand under $\theta(t)$. Given $P(\theta|H_{t-1}) = P(\theta(t)|H_{t-1})$ and because x^* is the optimal solution of LP(d) and $x(t)$ is the optimal solution of LP$(d(t))$, the solutions x^* and $x(t)$ are also identically distributed conditional on $\mathcal{H}_{t-1}$ (although again, this does not imply $x^* = x(t)$), namely

$$\mathbb{P}(x^* \mid \mathcal{H}_{t-1}) = \mathbb{P}(x(t) \mid \mathcal{H}_{t-1}).$$

(If there are multiple optimal solutions to either LP(d) or LP$(d(t))$, we assume the same tie-breaking rule is used.)

Therefore, the left-hand side of Eq (EC.3) can be decomposed using the law of iterated expectation as

$$\text{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k (x_k^* - x_k(t))\right]$$

$$
\begin{aligned}
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[\mathrm{E}[r_k x_k^* - r_k x_k(t) \mid \mathcal{H}_{t-1}]] \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[\mathrm{E}[r_k x_k^* - U_k(t)x_k^* + U_k(t)x_k(t) - r_k x_k(t) \mid \mathcal{H}_{t-1}]] \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[\mathrm{E}[r_k x_k^* - U_k(t)x_k^* \mid \mathcal{H}_{t-1}]] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[\mathrm{E}[U_k(t)x_k(t) - r_k x_k(t) \mid \mathcal{H}_{t-1}]] \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[r_k x_k^* - U_k(t)x_k^*] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[U_k(t)x_k(t) - r_k x_k(t)]. \qquad \text{(EC.4)}
\end{aligned}
$$

This decomposition is proposed by Russo and Van Roy (2014). Here, $U_k(t)$ is a deterministic function given $\mathcal{H}_{t-1}$, defined as follows. Let $N_k(t-1)$ be the number of times when price vector p_k is offered over the first $t-1$ periods, and let $\hat{d}_{ik}(t-1) = (\sum_{s=1}^{t-1} \mathbf{1}_{\{P(s)=p_k\}} D_i(s))/N_k(t-1)$ be the average demand of product i when price p_k is offered in the first $t-1$ periods. The function $U_k(t)$ is defined by

$$
U_k(t) = \sum_{i=1}^{N} p_{ik} \min\left(\hat{d}_{ik}(t-1) + \bar{d}_i \sqrt{\frac{\log\left(\frac{TK}{N_k(t-1)}\right)}{N_k(t-1)}}, \bar{d}_i \right). \qquad \text{(EC.5)}
$$

In the multi-armed bandit literature, $U_k(t)$ is known as an upper confidence bound (UCB), since it is an upper bound of r_k with high probability. Intuitively, if $U_k(t)$ is an upper bound of r_k, the first term of Eq (EC.4) $(r_k x_k^* - U_k(t)x_k^*)$ is non-positive, while the second term $(U_k(t)x_k(t) - r_k x_k(t))$ converges to zero as $N_k(t-1)$ increases. So both terms of Eq (EC.4) will eventually vanish.

More precisely, by Lemma EC.1 and EC.2, we have

$$
\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[r_k x_k^* - U_k(t)x_k^*] \le p_{\max} \cdot 6\sqrt{KT}; \quad \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}[U_k(t)x_k(t) - r_k x_k(t)] \le p_{\max} \cdot 12\sqrt{KT\log(K)}.
$$

This gives us the bound in Eq (EC.3).

Part II: Inventory Bound. Next, we show that

$$
\mathrm{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} a_{ij} D_i(t) - I_j \right)^+ \right] \le (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot 37\sqrt{KT\log K}, \quad \forall j \in [M]. \qquad \text{(EC.6)}
$$

We decompose the left hand side of Eq (EC.6) as

$$
\begin{aligned}
&\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N} a_{ij} D_i(t) - I_j \right)^+ \right] \\
=&\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N} \left(a_{ij} D_i(t) - \sum_{k=1}^{K} a_{ij} d_{ik} x_k(t) + \sum_{k=1}^{K} a_{ij} d_{ik} x_k(t) \right) - I_j \right)^+ \right] \\
\le&\mathrm{E}\left[\left| \sum_{t=1}^{T}\sum_{i=1}^{N} \left(a_{ij} D_i(t) - \sum_{k=1}^{K} a_{ij} d_{ik} x_k(t) \right) \right| \right] + \mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K} a_{ij} d_{ik} x_k(t) - I_j \right)^+ \right] \\
\le&\underbrace{\sum_{i=1}^{N} a_{ij} \mathrm{E}\left[\left| \sum_{t=1}^{T} \left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t) \right) \right| \right]}_{(\dagger)} + \underbrace{\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{k=1}^{K} b_{jk} x_k(t) - I_j \right)^+ \right]}_{(\ddagger)}.
\end{aligned}
$$

In the last inequality, we use the definition $b_{jk} = \sum_{i=1}^{N} a_{ij} d_{ik}$.

We first consider term (†). For any $i \in [N]$, using the fact that $\mathrm{E}[|X|]^2 \le \mathrm{E}[X^2]$ for any random variable X, we have

$$\begin{aligned}
&\mathrm{E}\left[\left|\sum_{t=1}^{T}\left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t)\right)\right|\right]^2 \\
\le &\mathrm{E}\left[\left(\sum_{t=1}^{T}\left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t)\right)\right)^2\right] \\
= &\mathrm{E}\left[\sum_{t=1}^{T}\left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t)\right)^2\right] + \mathrm{E}\left[2\sum_{1\le s<t\le T}\left(D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)\right)\left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t)\right)\right].
\end{aligned}$$

The second term in the last equation is equal to zero, because conditional on $(\mathcal{H}_{t-1}, d, \theta(s), \theta(t))$, the demand $D_i(t)$ has conditional expectation $\sum_{k=1}^{K} d_{ik} x_k(t)$, while $D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)$ is a constant. Thus, we have

$$\begin{aligned}
&\mathrm{E}\left[\left(D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)\right)\left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t)\right)\right] \\
= &\mathrm{E}\left[\left(D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)\right)\mathrm{E}\left[\left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t)\right) \mid \mathcal{H}_{t-1}, d, \theta(s), \theta(t)\right]\right] \\
= &\mathrm{E}\left[\left(D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)\right)\cdot 0\right] = 0.
\end{aligned}$$

Therefore, we have

$$(\dagger) \le \sum_{i=1}^{N} a_{ij} \mathrm{E}\left[\sum_{t=1}^{T}\left(D_i(t) - \sum_{k=1}^{K} d_{ik} x_k(t)\right)^2\right]^{\frac{1}{2}} \le \Big(\sum_{i=1}^{N} a_{ij} \bar{d}_i\Big) \cdot \sqrt{T}, \tag{EC.7}$$

where we use the assumption that $D_i(t)$ is upper bounded by $\bar{d}_i$ almost surely.

We now consider the term (‡). For the optimal solution of $\mathsf{LP}(d(t))$, it holds almost surely that

$$\sum_{k=1}^{K} b_{jk}(t) x_k(t) \le c_j = I_j / T.$$

Therefore, we have

$$\begin{aligned}
(\ddagger) = &\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{k=1}^{K} b_{jk} x_k(t) - \sum_{t=1}^{T} c_j\right)^+\right] \\
\le &\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{k=1}^{K} b_{jk} x_k(t) - \sum_{t=1}^{T}\sum_{k=1}^{K} b_{jk}(t) x_k(t)\right)^+\right] \\
= &\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{k=1}^{K} (b_{jk} - b_{jk}(t)) x_k(t)\right)^+\right] \\
\le &\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(b_{jk} - b_{jk}(t))^+ x_k(t)\right].
\end{aligned} \tag{EC.8}$$

In the last step, we use the fact that $(\alpha+\beta)^+ \le \alpha^+ + \beta^+$ for any $\alpha, \beta \in \mathbb{R}$.

Similar to Eq (EC.5) in Part (I), we define the upper confidence bound (UCB) function

$$U_{jk}(t) = \sum_{i=1}^{N} a_{ij} \min\left(\hat{d}_{ik}(t-1) + \bar{d}_i \sqrt{\frac{\log\left(\frac{TK}{N_k(t-1)}\right)}{N_k(t-1)}}, \bar{d}_i \right),$$

and the lower confidence bound (LCB) function

$$L_{jk}(t) = \sum_{i=1}^{N} a_{ij} \max\left(\hat{d}_{ik}(t-1) - \bar{d}_i \sqrt{\frac{\log\left(\frac{TK}{N_k(t-1)}\right)}{N_k(t-1)}}, 0 \right).$$

Since $U_{jk}(t) \geq L_{jk}(t)$, by Eq (EC.8) we have

$$\begin{aligned}
(\ddagger) \leq & \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(b_{jk} - U_{jk}(t) + U_{jk}(t) - L_{jk}(t) + L_{jk}(t) - b_{jk}(t))^+ x_k(t)\right] \\
\leq & \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(b_{jk} - U_{jk}(t))^+ x_k(t)\right] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(U_{jk}(t) - L_{jk}(t)) x_k(t)\right] \\
& + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(L_{jk}(t) - b_{jk}(t))^+ x_k(t)\right] \\
\leq & \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(b_{jk} - U_{jk}(t))^+\right] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(U_{jk}(t) - L_{jk}(t)) x_k(t)\right] \\
& + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(L_{jk}(t) - b_{jk}(t))^+\right].
\end{aligned}$$

In the second step, we again use the fact that $(\alpha+\beta+\gamma)^+ \leq \alpha^+ + \beta^+ + \gamma^+$ for any $\alpha, \beta, \gamma \in \mathbb{R}$. In the last step, we use the fact that $0 \leq x_k(t) \leq 1$.

By Lemma EC.1, we have

$$\sum_{k=1}^{K}\sum_{t=1}^{T} \mathrm{E}\left[(b_{jk} - U_{jk}(t))^+\right] \leq (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot 6\sqrt{KT}, \quad \sum_{k=1}^{K}\sum_{t=1}^{T} \mathrm{E}\left[(L_{jk}(t) - b_{jk})^+\right] \leq (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot 6\sqrt{KT}.$$

Because θ and $\theta(t)$ are identically distributed given $\mathcal{H}_{t-1}$, *i.e.*, $\mathbb{P}(\theta \mid \mathcal{H}_{t-1}) = \mathbb{P}(\theta(t) \mid \mathcal{H}_{t-1})$, we have $\mathbb{P}(b_{jk} \mid \mathcal{H}_{t-1}) = \mathbb{P}(b_{jk}(t) \mid \mathcal{H}_{t-1})$. Since $L_{jk}(t)$ is deterministic given $\mathcal{H}_{t-1}$, by the law of iterated expectation, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{t=1}^{T} \mathrm{E}\left[(L_{jk}(t) - b_{jk}(t))^+\right] &= \sum_{k=1}^{K}\sum_{t=1}^{T} \mathrm{E}\left[\mathrm{E}[(L_{jk}(t) - b_{jk}(t))^+ \mid \mathcal{H}_{t-1}]\right] \\
&= \sum_{k=1}^{K}\sum_{t=1}^{T} \mathrm{E}\left[\mathrm{E}[(L_{jk}(t) - b_{jk})^+ \mid \mathcal{H}_{t-1}]\right] \\
&= \sum_{k=1}^{K}\sum_{t=1}^{T} \mathrm{E}\left[(L_{jk}(t) - b_{jk})^+\right] \\
&\leq (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot 6\sqrt{KT}.
\end{aligned}$$

In addition, by Lemma EC.2, we have

$$\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(U_{jk}(t) - L_{jk}(t)) x_k(t)\right] \leq (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot 24\sqrt{KT\log K}.$$

Combining the above results, we have

$$(\ddagger) \le (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot (6+6+24\sqrt{\log K})\sqrt{KT} \le (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot 36\sqrt{KT\log K}. \tag{EC.9}$$

Adding both Eq (EC.7) and (EC.9), we complete the proof of Eq (EC.6). In the final step, combining Eq (EC.3) and (EC.6), we establish an upper bound of the Bayesian regret of TS-fixed as

$$\text{BayesRegret}(T) \le \left(18p_{\max} + 37\left(\sum_{j=1}^{M} p^j_{\max} \sum_{i=1}^{N} a_{ij}\bar{d}_i\right)\right) \cdot \sqrt{TK\log K}.$$

EC.2. Proof for TS-update

THEOREM 2. *The Bayesian regret of* TS-update *is bounded by*

$$\text{BayesRegret}(T) \le \left(18p_{\max} + 40\sum_{i=1}^{N}\sum_{j=1}^{M} p^j_{\max} a_{ij}\bar{d}_i\right)\sqrt{KT\log K} + p_{\max}M.$$

Proof. Let the mean demand under $\theta(t)$, the sampled parameter at period t, be $d(t) = \{d_{ik}(t)\}_{i\in[N],k\in[K]}$. We denote by $r_k(t) = \sum_{i=1}^{N} p_{ik}d_{ik}(t)$ and $b_{jk}(t) = \sum_{i=1}^{N} a_{ij}d_{ik}(t)$ the revenue and resource consumption under the demand sampled at period t, and denote by $c_j(t) = I_j(t-1)/(T-t+1)$ the inventory rate for resource j at the beginning of period t. Recall that $x(t)$ is the optimal solution of the following LP:

$$\begin{aligned}
\mathsf{LP}(d(t), c(t)): \quad \max_{x} \quad & \sum_{k=1}^{K} r_k(t)x_k \\
\text{subject to} \quad & \sum_{k=1}^{K} b_{jk}(t)x_k \le c_j(t), \quad \forall j \in [M] \\
& \sum_{k=1}^{K} x_k \le 1 \\
& x_k \ge 0, \quad \forall k \in [K].
\end{aligned}$$

If we assume unlimited inventory, the expected revenue of TS-update over the selling horizon is given by

$$\text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k \mathbf{1}_{\{P(t)=p_k\}}] = \text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k \text{E}[\mathbf{1}_{\{P(t)=p_k\}} \mid \theta, \theta(t)]] = \text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k x_k(t)]. \tag{EC.10}$$

To calculate the actual revenue, we should subtract from Eq (EC.10) the amount of revenue associated with lost sales. For TS-update, we claim that this amount is no more than $p_{\max}M$, where $p_{\max} = \max_{k\in[K]}\{\sum_{i=1}^{N} p_{ik}\bar{d}_i\}$ is the maximum revenue that can possibly be achieved in a single period and M is the number of resources. To show this, for any resource $j = 1, \ldots, M$, we define $T_j = \max\{t \mid c_j(t) > 0, t = 1, \ldots, T\}$. When $t < T_j$, the inventory level of resource j at the end of period t is positive. When $t > T_j$, we have $c_j(t) = 0$, so the optimal solution of TS-update, $x(t)$, will not use resource j in this period. Therefore, the case when TS-update consumes resource j over its inventory limit can only happen at period $t = T_j$. Since there are M resources in total, there can be at most M such periods where TS-update uses resources over the limit and incurs lost sales. The maximum revenue that can be achieved in these periods is $p_{\max}M$. Therefore, the actual revenue after accounting for lost sales is bounded by

$$\text{E}[\text{Revenue}(T)] \ge \text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k x_k(t)] - p_{\max}M. \tag{EC.11}$$

In Section 3.1.1, we have shown that

$$\text{BayesRegret}(T) \le T \cdot \text{E}[\sum_{k=1}^{K} r_k x_k^*] - \text{E}[\text{Revenue}(T)].$$

By Eq (EC.11), we have

$$\text{BayesRegret}(T) \le T \cdot \text{E}[\sum_{k=1}^{K} r_k x_k^*] - \text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k x_k(t)] + p_{\max} M. \tag{EC.12}$$

We consider another linear program where the sampled demands are replaced by true demands, while the inventory rate at the beginning of period t is still given by $c_j(t)$. For each price $k = 1, \ldots, K$, we denote by $r_k = \sum_{i=1}^{N} p_{ik} d_{ik}$ the mean revenue, and denote by $b_{jk} = \sum_{i=1}^{N} a_{ij} d_{ik}$ the expected consumption of resource j. We define the following linear program $\mathsf{LP}(d, c(t))$ as:

$$\begin{aligned}
\mathsf{LP}(d, c(t)): \quad \max_{x} \ & \sum_{k=1}^{K} r_k x_k \\
\text{subject to } & \sum_{k=1}^{K} b_{jk} x_k \le c_j(t), \quad \forall j \in [M] \\
& \sum_{k=1}^{K} x_k \le 1 \\
& x_k \ge 0, \quad \forall k \in [K].
\end{aligned}$$

We denote the optimal solution of $\mathsf{LP}(d, c(t))$ by $x^*(t)$.

By Eq (EC.12), we have

$$\begin{aligned}
\text{BayesRegret}(T) \le & \text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^* - r_k x_k(t))] + p_{\max} M \\
= & \text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^* - r_k x_k^*(t) + r_k x_k^*(t) - r_k x_k(t))] + p_{\max} M \\
= & \underbrace{\text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^*(t) - r_k x_k(t))]}_{\text{(I)}} + \underbrace{\text{E}[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^* - r_k x_k^*(t))]}_{\text{(II)}} + p_{\max} M.
\end{aligned}$$

To complete the proof, we show that both (I) and (II) are bounded by $O(\sqrt{T})$.

Part (I). First, we show that

$$\text{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^*(t) - r_k x_k(t))\right] \le p_{\max} \cdot 18\sqrt{KT \log K}, \tag{EC.13}$$

The proof for Eq (EC.15) is almost identical to the proof for Eq (EC.3) for TS-fixed. Let $\mathcal{H}_{t-1} = (P(1), D(1), \ldots, \theta(t-1), P(t-1), D(t-1))$ be the history available at the beginning of period t. In the TS-update algorithm, note that the values of $\theta(t)$ are sampled from the posterior distribution of θ conditional on history $\mathcal{H}_{t-1}$, *i.e.* $\mathbb{P}(\theta \mid \mathcal{H}_{t-1}) = \mathbb{P}(\theta(t) \mid \mathcal{H}_{t-1})$; see the explanation in EC.1. Because $x^*(t)$ is the optimal solution of $\mathsf{LP}(d, c(t))$ and $x(t)$ is the optimal solution of $\mathsf{LP}(d(t), c(t))$, and because $c(t)$ is deterministic given $\mathcal{H}_{t-1}$, the solutions x^* and $x(t)$ are identically distributed conditional on $\mathcal{H}_{t-1}$, namely

$$\mathbb{P}(x^*(t) \mid \mathcal{H}_{t-1}) = \mathbb{P}(x(t) \mid \mathcal{H}_{t-1}).$$

Therefore, using the law of iterated expectation, the left hand side of Eq (EC.15) can be decomposed as

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^*(t) - r_k x_k(t))\right] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[\mathrm{E}[r_k x_k^*(t) - r_k x_k(t) \mid \mathcal{H}_{t-1}]] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[\mathrm{E}[r_k x_k^*(t) - U_k(t)x_k^*(t) + U_k(t)x_k(t) - r_k x_k(t) \mid \mathcal{H}_{t-1}]] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[\mathrm{E}[r_k x_k^*(t) - U_k(t)x_k^*(t) \mid \mathcal{H}_{t-1}]] + \sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[\mathrm{E}[U_k(t)x_k(t) - r_k x_k(t) \mid \mathcal{H}_{t-1}]] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[r_k x_k^*(t) - U_k(t)x_k^*(t)] + \sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[U_k(t)x_k(t) - r_k x_k(t)]. \qquad \text{(EC.14)}
\end{aligned}$$

Here, $U_k(t)$ is a deterministic function given $\mathcal{H}_{t-1}$, defined by

$$U_k(t) = \sum_{i=1}^{N} p_{ik} \min\left(\hat{d}_{ik}(t-1) + \bar{d}_i\sqrt{\frac{\log\left(\frac{TK}{N_k(t-1)}\right)}{N_k(t-1)}}, \bar{d}_i\right).$$

We denote by $N_k(t-1)$ the number of times when price (vector) p_k is offered over the first $t-1$ periods, and denote by $\hat{d}_{ik}(t-1) = (\sum_{s=1}^{t-1}\mathbf{1}_{\{P(s)=p_k\}}D_i(s))/N_k(t-1)$ the average demand of product i when price p_k is offered in the first $t-1$ periods.

By Lemma EC.1, we have

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[r_k x_k^*(t) - U_k(t)x_k^*(t)] \le \sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}[(r_k - U_k(t))^+] \le p_{\max}\cdot 6\sqrt{KT}.$$

In addition, Lemma EC.2 shows that

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(U_k(t)x_k(t) - r_k x_k(t))\right] \le p_{\max}\cdot 12\sqrt{KT\log K}.$$

Substituting the above bounds in Eq (EC.14), we get Eq (EC.13).

Part (II). Next, we prove that

$$\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^* - r_k x_k^*(t))\right] \le \sum_{j=1}^{M}\left(p_{\max}^j\sum_{i=1}^{N}a_{ij}\bar{d}_i\right)\cdot 40\sqrt{KT\log K}. \qquad \text{(EC.15)}$$

Recall that by the definition of $\mathsf{LP}(d)$, $\sum_{k=1}^{K}(r_k x_k^*)$ is the optimal revenue per time period, assuming that the demand is deterministic and the quantity of resource j available at this period is c_j. Likewise, by the definition of $\mathsf{LP}(d, c(t))$, $\sum_{k=1}^{K}(r_k x_k^*(t))$ is the optimal revenue per time period, assuming the same deterministic demand, while the quantity of resource j is replaced by $c_j(t)$.

By definition, increasing resource j by one unit would increase the optimal value of $\mathsf{LP}(d, c(t))$ by at most $p_{\max}^j$. Therefore, for any $t = 1, \ldots, K$, we have

$$\sum_{k=1}^{K}(r_k x_k^* - r_k x_k^*(t)) \le \sum_{j=1}^{M}p_{\max}^j(c_j - c_j(t))^+. \qquad \text{(EC.16)}$$

It holds that $c_j = \frac{I_j}{T}$ and $c_j(t) = \frac{I_j(t-1)}{T-t+1}$. In addition, we have

$$I_j(s-1) - I_j(s) \le \sum_{i=1}^{N} a_{ij} D_i(s), \quad \forall 1 \le s \le t. \tag{EC.17}$$

Eq (EC.17) holds as an inequality, not as an equality, because there may be lost sales in period s. By Eq (EC.17), we have

$$\begin{aligned}
\mathrm{E}\left[c_j - c_j(t)\right] =& \mathrm{E}\left[\sum_{s=1}^{t-1}(c_j(s) - c_j(s+1))\right] \\
=& \mathrm{E}\left[\sum_{s=1}^{t-1}\left(\frac{I_j(s-1)}{T-s+1} - \frac{I_j(s)}{T-s}\right)\right] \\
=& \mathrm{E}\left[\sum_{s=1}^{t-1}\left(\frac{I_j(s-1)}{T-s} - \frac{I_j(s)}{T-s} - \frac{I_j(s-1)}{(T-s+1)(T-s)}\right)\right] \\
\le& \mathrm{E}\left[\sum_{s=1}^{t-1}\left(\frac{\sum_{i=1}^{N} a_{ij} D_i(s)}{T-s} - \frac{c_j(s)}{T-s}\right)\right] \\
=& \mathrm{E}\left[\sum_{s=1}^{t-1}\left(\frac{\sum_{i=1}^{N} a_{ij} D_i(s)}{T-s} - \frac{\sum_{k=1}^{K} b_{jk} x_k(s)}{T-s}\right)\right] \\
&+ \mathrm{E}\left[\sum_{s=1}^{t-1}\left(\frac{\sum_{k=1}^{K} b_{jk} x_k(s)}{T-s} - \frac{c_j(s)}{T-s}\right)\right].
\end{aligned}$$

Applying function $(x)^+$ to both sides and using the fact that $(\alpha+\beta)^+ \le \alpha^+ + \beta^+$ for any $\alpha, \beta \in \mathbb{R}$, we have

$$\begin{aligned}
\mathrm{E}\left[(c_j - c_j(t))^+\right] \le& \mathrm{E}\left[\left(\sum_{s=1}^{t-1} \frac{\sum_{i=1}^{N} a_{ij} D_i(s) - \sum_{k=1}^{K} b_{jk} x_k(s)}{T-s}\right)^+\right] && \text{(EC.18)} \\
&+ \mathrm{E}\left[\left(\sum_{s=1}^{t-1} \frac{\sum_{k=1}^{K} b_{jk} x_k(s) - c_j(s)}{T-s}\right)^+\right] \\
\le& \underbrace{\mathrm{E}\left[\sum_{i=1}^{N} a_{ij}\left(\sum_{s=1}^{t-1} \frac{D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)}{T-s}\right)^+\right]}_{(\dagger)} && \text{(EC.19)} \\
&+ \mathrm{E}\left[\left(\sum_{s=1}^{t-1} \frac{\sum_{k=1}^{K} b_{jk} x_k(s) - \sum_{k=1}^{K} b_{jk}(s) x_k(s)}{T-s}\right)^+\right], && \text{(EC.20)}
\end{aligned}$$

where the last step uses the definition $b_{jk} = \sum_{i=1}^{N} a_{ij} d_{ik}$ and the fact $\sum_{k=1}^{K} b_{jk}(s) x_k(s) \le c_j(s), \forall 1 \le s \le t-1$.

We first focus on the term (†). For any $i = 1, \ldots, N$, using the fact that $\mathrm{E}[|X|]^2 \le \mathrm{E}[X^2]$ for any random variable X, we have

$$\begin{aligned}
&\mathrm{E}\left[\left|\sum_{s=1}^{t-1} \frac{D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)}{T-s}\right|\right]^2 \\
\le& \mathrm{E}\left[\left(\sum_{s=1}^{t-1} \frac{D_i(s) - \sum_{k=1}^{K} d_{ik} x_k(s)}{T-s}\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
=&\mathrm{E}\left[\sum_{s=1}^{t-1}\left(\frac{D_i(s)-\sum_{k=1}^K d_{ik}x_k(s)}{T-s}\right)^2\right]\\
&+\mathrm{E}\left[2\sum_{1\le s<s'\le t-1}\left(\frac{D_i(s)-\sum_{k=1}^K d_{ik}x_k(s)}{T-s}\right)\left(\frac{D_i(s')-\sum_{k=1}^K d_{ik}x_k(s')}{T-s'}\right)\right].
\end{aligned}
$$

The second term in the last equation is equal to zero, because conditional on $(\mathcal{H}_{s'-1}, d, \theta(s), \theta(s'))$, the demand $D_i(s'-1)$ has conditional mean $\sum_{k=1}^K d_{ik}x_k(s'-1)$, while $D_i(s)-\sum_{k=1}^K d_{ik}x_k(s)$ is a constant. Thus, we have

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{D_i(s)-\sum_{k=1}^K d_{ik}x_k(s)}{T-s}\right)\left(\frac{D_i(s')-\sum_{k=1}^K d_{ik}x_k(s')}{T-s'}\right)\right]\\
=&\mathrm{E}\left[\left(\frac{D_i(s)-\sum_{k=1}^K d_{ik}x_k(s)}{T-s}\right)\mathrm{E}\left[\left(\frac{D_i(s')-\sum_{k=1}^K d_{ik}x_k(s')}{T-s'}\right)\mid \mathcal{H}_{s'-1}, d, \theta(s), \theta(s')\right]\right]\\
=&\mathrm{E}\left[\left(\frac{D_i(s)-\sum_{k=1}^K d_{ik}x_k(s)}{T-s}\right)\cdot 0\right]=0.
\end{aligned}
$$

Therefore, we have

$$
(\dagger)\le\sum_{i=1}^N a_{ij}\mathrm{E}\left[\sum_{s=1}^{t-1}\left(\frac{D_i(s)-\sum_{k=1}^K d_{ik}x_k(s)}{T-s}\right)^2\right]^{\frac{1}{2}}\le(\sum_{i=1}^N a_{ij}\bar{d}_i)\cdot\left(\sum_{s=1}^{t-1}\frac{1}{(T-s)^2}\right)^{\frac{1}{2}}. \tag{EC.21}
$$

In sum, by Eq (EC.16), Eq (EC.20) and (EC.21), we have

$$
\begin{aligned}
\mathrm{E}\left[\sum_{t=1}^T\sum_{k=1}^K(r_kx_k^*-r_kx_k^*(t))\right]\le&\mathrm{E}\left[\sum_{t=1}^T\sum_{j=1}^M p^j_{\max}(c_j-c_j(t))^+\right]\\
\le&\sum_{j=1}^M\left(p^j_{\max}\sum_{i=1}^N a_{ij}\bar{d}_i\right)\cdot\sum_{t=1}^T\left(\sum_{s=1}^{t-1}\frac{1}{(T-s)^2}\right)^{\frac{1}{2}}\\
&+\sum_{j=1}^M p^j_{\max}\sum_{t=1}^T\sum_{s=1}^{t-1}\frac{1}{T-s}\mathrm{E}\left[\left(\sum_{k=1}^K b_{jk}x_k(s)-\sum_{k=1}^K b_{jk}(s)x_k(s)\right)^+\right]\\
\le&\sum_{j=1}^M\left(p^j_{\max}\sum_{i=1}^N a_{ij}\bar{d}_i\right)\cdot(2\sqrt{T}+\sqrt{2}) \qquad \text{(EC.22)}\\
&+\sum_{j=1}^M p^j_{\max}\sum_{s=1}^T\sum_{t=s+1}^T\frac{1}{T-s}\mathrm{E}\left[\left(\sum_{k=1}^K b_{jk}x_k(s)-\sum_{k=1}^K b_{jk}(s)x_k(s)\right)^+\right] \qquad \text{(EC.23)}\\
\le&\sum_{j=1}^M\left(p^j_{\max}\sum_{i=1}^N a_{ij}\bar{d}_i\right)\cdot(2\sqrt{T}+\sqrt{2})\\
&+\sum_{j=1}^M p^j_{\max}\sum_{s=1}^T\mathrm{E}\left[\left(\sum_{k=1}^K b_{jk}x_k(s)-\sum_{k=1}^K b_{jk}(s)x_k(s)\right)^+\right].
\end{aligned}
$$

In step (EC.23), we changed the order of sums. In step (EC.22), we use the fact that

$$
\begin{aligned}
&\sum_{t=1}^T\left(\sum_{s=1}^{t-1}\frac{1}{(T-s)^2}\right)^{\frac{1}{2}}\\
\le&\sum_{t=1}^{T-1}\left(\int_1^t\frac{1}{(T-s)^2}ds\right)^{\frac{1}{2}}+\left(\int_1^{T-1}\frac{1}{(T-s)^2}ds+1\right)^{\frac{1}{2}}
\end{aligned}
$$

$$\leq \sum_{t=1}^{T-1} \left(\frac{1}{T-t}\right)^{\frac{1}{2}} + \left(1 - \frac{1}{T-1} + 1\right)^{\frac{1}{2}}$$
$$\leq 2\sqrt{T} + \sqrt{2}.$$

Using the same proof for Eq (EC.8) and Eq (EC.9) for TS-fixed, we have

$$\sum_{s=1}^{T} \mathrm{E}\left[\left(\sum_{k=1}^{K} b_{jk} x_k(s) - \sum_{k=1}^{K} b_{jk}(s) x_k(s)\right)^+\right] \leq (\sum_{i=1}^{N} a_{ij}\bar{d}_i) \cdot 36\sqrt{KT\log K},$$

and therefore

$$\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}(r_k x_k^* - r_k x_k^*(t))\right] \leq \sum_{j=1}^{M}\left(p_{\max}^j \sum_{i=1}^{N} a_{ij}\bar{d}_i\right) \cdot (\sqrt{2} + 2\sqrt{T} + 36\sqrt{KT\log K})$$

which completes the proof for Eq (EC.15).

In the final step, combining Eq (EC.13) and (EC.15), we establish an upper bound of the Bayesian regret of TS-update:

$$\mathrm{BayesRegret}(T) \leq \left(18 p_{\max} + 40\left(\sum_{j=1}^{M} p_{\max}^j \sum_{i=1}^{N} a_{ij}\bar{d}_i\right)\right) \cdot \sqrt{TK\log K} + p_{\max} M.$$

EC.3. Proof for TS-linear

THEOREM 3. *The Bayesian regret of TS-linear is bounded by*

$$\mathrm{BayesRegret}(T) \leq O(N^2 \log T\sqrt{T}).$$

Proof. First, for any price vector $p \in \mathcal{P}$ and parameter $\theta' = (\alpha', B') \in \Theta$, we define the mean demand function $d : \mathcal{P} \times \Theta \to \mathbb{R}^N$ as $d(p, \theta') := \alpha' + B'p$. We use $d_i(p, \theta')$ to denote the i-th element of $d(p, \theta')$.

Let p^* be the optimal solution of the following QP:

$$\begin{aligned} \mathsf{QP}(\theta): \quad \max_{p} \quad & p^T d(p, \theta) \\ \text{subject to} \quad & A^T d(p, \theta) \leq c \\ & p \in \mathcal{P}. \end{aligned}$$

Similar to the discrete price case, the optimal value of $\mathsf{QP}(\theta)$ is an upper bound for the expected revenue of the full information case (see, *e.g.*, Chen et al. (2014)). Therefore, similar to the proof of TS-fixed, the Bayesian regret of TS-linear is bounded by

$$\mathrm{BayesRegret}(T) \leq \underbrace{\mathrm{E}\left[\sum_{t=1}^{T}(p^* \cdot d(p^*, \theta) - P_t \cdot d(P_t, \theta))\right]}_{\text{(I)}} + \underbrace{\sum_{j=1}^{M} p_{\max}^j \mathrm{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} a_{ij} D_i(t) - I_j\right)^+\right]}_{\text{(II)}}. \quad \text{(EC.24)}$$

(To simplify notation, we replace $P(t)$ and $\theta(t)$ with P_t and θ_t throughout this proof.)

Part (I) is the contribution to regret assuming unlimited inventory, and part (II) is the contribution to regret by accounting for lost sales. Here, $p_{\max}^j$ is the maximum revenue that can be achieved by adding one unit of resource j, defined by $p_{\max}^j := \max_{i:a_{ij}\neq 0, p\in\mathcal{P}}(p_i/a_{ij})$.

To complete the proof, we show that both (I) and (II) are bounded.

For part (I), since price set $\mathcal{P}$ and parameter Θ are bounded and the demand noise is sub-Gaussian, we can immediately apply the result of Russo and Van Roy (2014), Proposition 3, to get

$$\mathrm{E}\left[\sum_{t=1}^{T}(p^*\cdot d(p^*,\theta)-P_t\cdot d(P_t,\theta))\right]\le O(N^2\log T\sqrt{T}). \tag{EC.25}$$

For part (II), since $\sum_{i=1}^{N}a_{ij}d_i(P_t,\theta_t)\le c_j$ for all j, using the fact that $(\alpha+\beta)^+\le\alpha^++\beta^+$ for any $\alpha,\beta\in\mathbb{R}$, we have

$$\begin{aligned}
&\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}D_i(t)-I_j\right)^+\right]\\
\le&\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}\left(D_i(t)-d_i(P_t,\theta)\right)\right)^+\right]+\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}d_i(P_t,\theta)-c_jT\right)^+\right]\\
\le&\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}\left(D_i(t)-d_i(P_t,\theta)\right)\right)^+\right]+\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}(d_i(P_t,\theta)-d_i(P_t,\theta_t))\right)^+\right]\\
\le&\underbrace{\sum_{i=1}^{N}a_{ij}\mathrm{E}\left[\left|\sum_{t=1}^{T}\left(D_i(t)-d_i(P_t,\theta)\right)\right|\right]}_{(\dagger)}+\underbrace{\sum_{i=1}^{N}a_{ij}\mathrm{E}\left[\sum_{t=1}^{T}(d_i(P_t,\theta)-d_i(P_t,\theta_t))^+\right]}_{(\ddagger)}.
\end{aligned}$$

By the same proof for Eq (EC.7) for TS-fixed, we have

$$(\dagger)\le\sum_{i=1}^{N}a_{ij}\mathrm{E}\left[\sum_{t=1}^{T}\left(D_i(t)-d_i(P_t,\theta)\right)^2\right]^{\frac{1}{2}}=O(\sqrt{T}).$$

To bound (‡), we follow the steps in Russo and Van Roy (2014) to define upper confidence bound (UCB) and lower confidence bound (LCB) functions for the linear demand model. Let $\hat{\theta}_t^{LS}$ be the least squares estimator of θ given $\mathcal{H}_{t-1}$. We define an ellipsoidal confidence set

$$\Theta_t:=\{\theta'\in\mathbb{R}^N\times\mathbb{R}^{N^2}:\sum_{s=1}^{t-1}\left(d_i(P_s,\theta')-d_i(P_s,\hat{\theta}_t^{LS})\right)^2\le\beta_t\},$$

where β_t is a parameter defined in Russo and Van Roy (2014), Eq (8), where $\beta=O(N^2\log(t/T))$. We then define the UCB and LCB functions for all $i\in[N]$ as

$$U_i(p,t):=\max_{\theta'\in\Theta_t}d_i(p,\theta'),\quad L_i(p,t):=\min_{\theta'\in\Theta_t}d_i(p,\theta').$$

We decompose part (‡) as

$$\begin{aligned}
&\sum_{t=1}^{T}\mathrm{E}\left[(d_i(P_t,\theta)-d_i(P_t,\theta_t))^+\right]\\
\le&\sum_{t=1}^{T}\mathrm{E}\left[(d_i(P_t,\theta)-U_i(P_t,t))^+\right]+\sum_{t=1}^{T}\mathrm{E}\left[U_i(P_t,t)-L_i(P_t,t)\right]+\sum_{t=1}^{T}\mathrm{E}\left[(L_i(P_t,t)-d_i(P_t,\theta_t))^+\right]
\end{aligned}$$

Russo and Van Roy (2014), Proposition 6, shows that

$$\mathbb{P}\left(L_i(p,t)\le d_i(p,\theta)\le U_i(p,t),\forall t\in[T],\forall p\in\mathcal{P}\right)\ge1-\frac{1}{T}.$$

Conditional on $\mathcal{H}_{t-1}$, parameters θ and θ_t are identically distributed, while $U_i(p,t)$ and $L_i(p,t)$ are deterministic. Therefore, by the law of iterated expectation, we have

$$\begin{aligned}
&\mathbb{P}\left(L_i(p,t) \le d_i(p,\theta_t) \le U_i(p,t), \forall t \in [T], \forall p \in \mathcal{P}\right)\\
=&\mathrm{E}\left[\mathbf{1}\left(L_i(p,t) \le d_i(p,\theta_t) \le U_i(p,t), \forall t \in [T], \forall p \in \mathcal{P}\right)\right]\\
=&\mathrm{E}\left[\mathrm{E}\left[\mathbf{1}\left(L_i(p,t) \le d_i(p,\theta_t) \le U_i(p,t), \forall t \in [T], \forall p \in \mathcal{P}\right) \mid \mathcal{H}_{t-1}\right]\right]\\
=&\mathrm{E}\left[\mathrm{E}\left[\mathbf{1}\left(L_i(p,t) \le d_i(p,\theta) \le U_i(p,t), \forall t \in [T], \forall p \in \mathcal{P}\right) \mid \mathcal{H}_{t-1}\right]\right]\\
=&\mathrm{E}\left[\mathbf{1}\left(L_i(p,t) \le d_i(p,\theta) \le U_i(p,t), \forall t \in [T], \forall p \in \mathcal{P}\right)\right]\\
\ge&1-\frac{1}{T}.
\end{aligned}$$

This implies that

$$\sum_{t=1}^{T}\mathrm{E}\left[(d_i(P_t,\theta)-U_i(P_t,t))^+\right]=O(1), \quad \sum_{t=1}^{T}\mathrm{E}\left[(L_i(P_t,t)-d_i(P_t,\theta_t))^+\right]=O(1). \tag{EC.26}$$

Moreover, Russo and Van Roy (2014), Lemma 5, shows that

$$\sum_{t=1}^{T}\mathrm{E}\left[U_i(P_t,t)-L_i(P_t,t)\right]=O(N^2\log T\sqrt{T}). \tag{EC.27}$$

Adding Eq (EC.26) and (EC.27), we get

$$\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}D_i(t)-I_j\right)^+\right]\le(\dagger)+(\ddagger)=O(\sqrt{T})+O(N^2\log T\sqrt{T})+O(1)=O(N^2\log T\sqrt{T}).$$

Combining this result with Eq (EC.24) and (EC.25), we have

$$\mathrm{BayesRegret}(T)\le O(N^2\log T\sqrt{T})+O(N^2\log T\sqrt{T})=O(N^2\log T\sqrt{T}).$$

EC.4. Proof for TS-contextual

THEOREM 4. *The Bayesian regret of TS-contextual is bounded by*

$$\mathrm{BayesRegret}(T)\le\left(18p_{\max}+38\left(\sum_{j=1}^{M}p^j_{\max}\sum_{i=1}^{N}a_{ij}\bar{d}_i\right)\right)\cdot\sqrt{|\mathcal{X}|TK\log K}.$$

Proof. For each price $k\in[K]$ and each context $\xi\in\mathcal{X}$, we denote by $r_{\xi,k}=\sum_{i=1}^{N}p_{ik}d_{ik}(\xi\mid\theta)$ the mean revenue. Let x^* be the optimal solution of $\mathsf{LP}(\xi\mid\theta)$. It has been shown that the optimal value of $\mathsf{LP}(\xi\mid\theta)$ multiplied by T, or

$$\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}r_{\xi(t),k}x^*_{\xi(t),k}\right],$$

is an upper bound of the expected revenue with known demand information (Badanidiyuru et al. 2014).

Recall that $x(t)$ is the optimal solution of $\mathsf{LP}(\xi\mid\theta(t))$. The Bayesian regret is bounded by

$$\mathrm{BayesRegret}(T)\le\underbrace{\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}r_{\xi(t),k}(x^*_{\xi(t),k}-x_{\xi(t),k}(t))\right]}_{\text{(I)}}+\underbrace{\sum_{j=1}^{M}p^j_{\max}\mathrm{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T}a_{ij}D_i(t)-I_j\right)^+\right]}_{\text{(II)}}.$$

We will bound parts (I) and (II) separately.

Part (I). Let $\mathcal{H}_{t-1} = (\xi(1), P(1), D(1), \ldots, \xi(t-1), P(t-1), D(t-1))$ be the history available at the beginning of period t. In the TS-contextual algorithm, the parameter $\theta(t)$ is sampled from the posterior distribution given history $\mathcal{H}_{t-1}$, *i.e.* $\mathbb{P}(\theta \mid \mathcal{H}_{t-1}) = \mathbb{P}(\theta(t) \mid \mathcal{H}_{t-1})$. Since we assume $\xi(t)$ is sampled i.i.d. from a known distribution, it is independent of θ, $\theta(t)$, and $\mathcal{H}_{t-1}$, so $\mathbb{P}(\theta \mid \mathcal{H}_{t-1}, \xi(t)) = \mathbb{P}(\theta(t) \mid \mathcal{H}_{t-1}, \xi(t))$.

Because x^* is the optimal solution of $\mathsf{LP}(\xi \mid \theta)$ and $x(t)$ is the optimal solution of $\mathsf{LP}(\xi \mid \theta(t))$, the solutions x^* and $x(t)$ are also identically distributed conditional on $\mathcal{H}_{t-1}, \xi(t)$, namely

$$\mathbb{P}(x^* \mid \mathcal{H}_{t-1}, \xi(t)) = \mathbb{P}(x(t) \mid \mathcal{H}_{t-1}, \xi(t)).$$

Therefore, by the law of iterated expectation, the left hand side of Eq (EC.28) can be decomposed as

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_{\xi(t),k}(x^*_{\xi(t),k} - x_{\xi(t),k}(t))\right] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[\mathrm{E}[r_{\xi(t),k}x^*_{\xi(t),k} - r_{\xi(t),k}x_{\xi(t),k}(t) \mid \mathcal{H}_{t-1}, \xi(t)]\right] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[\mathrm{E}[r_{\xi(t),k}x^*_{\xi(t),k} - U_{\xi(t),k}(t)x^*_{\xi(t),k} + U_{\xi(t),k}(t)x_{\xi(t),k}(t) - r_{\xi(t),k}x_{\xi(t),k}(t) \mid \mathcal{H}_{t-1}, \xi(t)]\right] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[\mathrm{E}[r_{\xi(t),k}x^*_{\xi(t),k} - U_{\xi(t),k}(t)x^*_{\xi(t),k} \mid \mathcal{H}_{t-1}, \xi(t)]\right] \\
&+\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[\mathrm{E}[U_{\xi(t),k}(t)x_{\xi(t),k}(t) - r_{\xi(t),k}x_{\xi(t),k}(t) \mid \mathcal{H}_{t-1}, \xi(t)]\right] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[r_{\xi(t),k}x^*_{\xi(t),k} - U_{\xi(t),k}(t)x^*_{\xi(t),k}\right] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[U_{\xi(t),k}(t)x_{\xi(t),k}(t) - r_{\xi(t),k}x_{\xi(t),k}(t)\right] \\
=&\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[\sum_{\xi\in\mathcal{X}} \mathrm{E}[r_{\xi,k}x^*_{\xi,k} - U_{\xi,k}(t)x^*_{\xi,k} \mid \xi(t) = \xi]\mathbb{P}(\xi(t) = \xi)\right] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[U_{\xi(t),k}(t)x_{\xi(t),k}(t) - r_{\xi(t),k}x_{\xi(t),k}(t)\right]
\end{aligned}$$

Here, $U_{\xi,k}(t)$ is a deterministic function given $\mathcal{H}_{t-1}$, defined as

$$U_{\xi,k}(t) = \sum_{i=1}^{N} p_{ik} \min\left(\hat{d}_{ik\xi}(t-1) + \bar{d}_i\sqrt{\frac{\log_+\left(\frac{TK}{N_{\xi,k}(t-1)|\mathcal{X}|}\right)}{N_{\xi,k}(t-1)}}, \bar{d}_i\right),$$

where $\log_+(x) = \log(x)\mathbf{1}_{\{x\geq 1\}}$. $N_{\xi,k}(t-1)$ denotes the number of times when price (vector) p_k is offered with context ξ over the first $t-1$ periods, and $\hat{d}_{ik\xi}(t-1)$ is the average demand of product i when price p_k is offered with context ξ in the first $t-1$ periods.

By Lemma EC.3, for all $t \in [T]$, $k \in [K]$ and $\xi \in \mathcal{X}$, since $0 \leq x^*_{\xi,k} \leq 1$, we have

$$\mathrm{E}[r_{\xi,k}x^*_{\xi,k} - U_{\xi,k}(t)x^*_{\xi,k}] \leq \mathrm{E}[(r_{\xi,k} - U_{\xi,k}(t))^+] \leq p_{\max}\cdot\frac{6\sqrt{|\mathcal{X}|}}{\sqrt{KT}},$$

where $p_{\max} = \max_{k\in[K]}\{\sum_{i=1}^{N} p_{ik}\bar{d}_i\}$ is the maximum revenue that can be achieved in a single period. Since $\xi(t)$ is independent of θ and $\mathcal{H}_{t-1}$, we have

$$\mathrm{E}[r_{\xi,k}x^*_{\xi,k} - U_{\xi,k}(t)x^*_{\xi,k} \mid \xi(t)] = \mathrm{E}[r_{\xi,k}x^*_{\xi,k} - U_{\xi,k}(t)x^*_{\xi,k}] \leq p_{\max}\cdot\frac{6\sqrt{|\mathcal{X}|}}{\sqrt{KT}}.$$

Therefore,

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[\sum_{\xi\in\mathcal{X}}\mathrm{E}[r_{\xi,k}x^*_{\xi,k}-U_{\xi,k}(t)x^*_{\xi,k}\mid\xi(t)=\xi]\mathbb{P}(\xi(t)=\xi)\right]\le p_{\max}\cdot 6\sqrt{|\mathcal{X}|KT}.$$

By Lemma EC.4, we have

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[U_{\xi(t),k}(t)x_{\xi(t),k}(t)-r_{\xi(t),k}x_{\xi(t),k}(t)\right]\le p_{\max}\cdot 12\sqrt{|\mathcal{X}|KT\log K}.$$

This proves that

$$\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}r_{\xi(t),k}(x^*_{\xi(t),k}-x_{\xi(t),k}(t))\right]\le p_{\max}\cdot 18\sqrt{|\mathcal{X}|KT\log(K)}. \tag{EC.28}$$

Part II. We now consider part (II) in the regret bound. According to the definition of $\mathsf{LP}(\xi\mid\theta(t))$, it holds almost surely that

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\mathrm{E}_\xi[a_{ij}d_{ik}(\xi\mid\theta(t))x_{\xi,k}(t)]\le c_j=I_j/T,$$

where $\mathrm{E}_\xi[\cdot]$ is the expectation operator assuming ξ follows the distribution of features. Therefore,

$$(\mathrm{II})=\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}D_i(t)-I_j\right)^+\right]\le\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}\left(D_i(t)-\sum_{k=1}^{K}\mathrm{E}_\xi[d_{ik}(\xi\mid\theta(t))x_{\xi,k}(t)]\right)\right)^+\right]$$

We use the following decomposition:

$$\begin{aligned}
&\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}\left(D_i(t)-\sum_{k=1}^{K}\mathrm{E}_\xi[d_{ik}(\xi\mid\theta(t))x_{\xi,k}(t)]\right)\right)^+\right]\\
=&\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}\left(D_i(t)-\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta)x_{\xi,k}(t)\right)\right)^+\right] \quad (\dagger)\\
&+\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}\left(\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta)x_{\xi(t),k}(t)-\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta(t))x_{\xi(t),k}(t)\right)\right)^+\right] \quad (\ddagger)\\
&+\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{i=1}^{N}a_{ij}\left(\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta(t))x_{\xi(t),k}(t)-\sum_{k=1}^{K}\mathrm{E}_\xi[d_{ik}(\xi\mid\theta(t))x_{\xi,k}(t)]\right)\right)^+\right]. \quad (\dagger')
\end{aligned}$$

We first consider term (†). For any $i=1,\dots,N$, using the fact that $\mathrm{E}[|X|]^2\le\mathrm{E}[X^2]$ for any random variable X, we have

$$\begin{aligned}
&\mathrm{E}\left[\left|\sum_{t=1}^{T}\left(D_i(t)-\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta)x_{\xi,k}(t)\right)\right|\right]^2\\
\le&\mathrm{E}\left[\left(\sum_{t=1}^{T}\left(D_i(t)-\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta)x_{\xi,k}(t)\right)\right)^2\right]\\
=&\mathrm{E}\left[\sum_{t=1}^{T}\left(D_i(t)-\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta)x_{\xi,k}(t)\right)^2\right]\le(\bar{d}_i)^2T.
\end{aligned}$$

The equality in the last line holds because conditional on $(\theta,\xi(t),x(t))$, demand $D_i(t)$ has mean $\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta)x_{\xi,k}(t)$ and is uncorrelated with history $\mathcal{H}_{t-1}$. Therefore, we have

$$(\dagger)\le\sum_{i=1}^{N}a_{ij}\mathrm{E}\left[\sum_{t=1}^{T}\left(D_i(t)-\sum_{k=1}^{K}d_{ik}(\xi(t)\mid\theta)x_{\xi,k}(t)\right)^2\right]^{\frac{1}{2}}\le(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot\sqrt{T}. \tag{EC.29}$$

We then consider term (†′). Again using the fact that $\mathrm{E}[|X|]^2 \le \mathrm{E}[X^2]$ for any random variable X, for any $i = 1, \ldots, N$, we have

$$\begin{aligned}
&\mathrm{E}\left[\left|\sum_{t=1}^{T}\left(\sum_{k=1}^{K} d_{ik}(\xi(t) \mid \theta(t)) x_{\xi(t),k}(t) - \sum_{k=1}^{K} \mathrm{E}_{\xi}[d_{ik}(\xi \mid \theta(t)) x_{\xi,k}(t)]\right)\right|\right]^2 \\
\le &\mathrm{E}\left[\left(\sum_{t=1}^{T}\left(\sum_{k=1}^{K} d_{ik}(\xi(t) \mid \theta(t)) x_{\xi(t),k}(t) - \sum_{k=1}^{K} \mathrm{E}_{\xi}[d_{ik}(\xi \mid \theta(t)) x_{\xi,k}(t)]\right)\right)^2\right] \\
= &\mathrm{E}\left[\sum_{t=1}^{T}\left(\sum_{k=1}^{K} d_{ik}(\xi(t) \mid \theta(t)) x_{\xi(t),k}(t) - \sum_{k=1}^{K} \mathrm{E}_{\xi}[d_{ik}(\xi \mid \theta(t)) x_{\xi,k}(t)]\right)^2\right] \le (\bar{d}_i)^2 T.
\end{aligned}$$

The equality in the last line holds because conditional on $(\theta(t), x(t))$, $\sum_{k=1}^{K} d_{ik}(\xi(t) \mid \theta(t)) x_{\xi(t),k}(t)$ has mean $\sum_{k=1}^{K} \mathrm{E}_{\xi}[d_{ik}(\xi \mid \theta(t)) x_{\xi,k}(t)]$ and is uncorrelated with $\mathcal{H}_{t-1}$. Therefore, we have

$$(\dagger') \le \sum_{i=1}^{N} a_{ij} \mathrm{E}\left[\sum_{t=1}^{T}\left(\sum_{k=1}^{K} d_{ik}(\xi(t) \mid \theta(t)) x_{\xi(t),k}(t) - \sum_{k=1}^{K} \mathrm{E}_{\xi}[d_{ik}(\xi \mid \theta(t)) x_{\xi,k}(t)]\right)^2\right]^{\frac{1}{2}} \le (\sum_{i=1}^{N} a_{ij} \bar{d}_i) \cdot \sqrt{T}. \tag{EC.30}$$

We finally consider term (‡). We denote by $b_{jk\xi} = \sum_{i=1}^{N} a_{ij} d_{ik}(\xi \mid \theta)$ and $b_{jk\xi}(t) = \sum_{i=1}^{N} a_{ij} d_{ik}(\xi \mid \theta(t))$. Using the fact that $(\alpha + \beta)^+ \le \alpha^+ + \beta^+$ for any $\alpha, \beta \in \mathbb{R}$, we have

$$\begin{aligned}
(\ddagger) = &\mathrm{E}\left[\left(\sum_{t=1}^{T}\sum_{k=1}^{K} b_{jk\xi(t)} x_{\xi(t),k}(t) - \sum_{t=1}^{T}\sum_{k=1}^{K} b_{jk\xi(t)}(t) x_{\xi(t),k}(t)\right)^+\right] \\
\le &\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(b_{jk\xi(t)} - b_{jk\xi(t)}(t))^+ x_{\xi(t),k}(t)\right].
\end{aligned} \tag{EC.31}$$

For all $k \in [K]$, $\xi \in \mathcal{X}$, let $N_{\xi,k}(t-1)$ be the number of times that price vector p_k is offered with context ξ over the first $t-1$ periods, and let $\hat{d}_{ik\xi}(t-1) = (\sum_{s=1}^{t-1} \mathbf{1}_{\{P(s)=p_k, \xi(s)=\xi\}} D_i(s)) / N_{\xi,k}(t-1)$ be the average demand of product i when price p_k is offered with context ξ in the first $t-1$ periods. Define the upper confidence bound (UCB) function as

$$U_{jk\xi}(t) = \sum_{i=1}^{N} a_{ij} \min\left(\hat{d}_{ik\xi}(t-1) + \bar{d}_i \sqrt{\frac{\log_+\left(\frac{TK}{N_{\xi,k}(t-1)|\mathcal{X}|}\right)}{N_{\xi,k}(t-1)}}, \bar{d}_i\right),$$

and the lower confidence bound (LCB) function as

$$L_{jk\xi}(t) = \sum_{i=1}^{N} a_{ij} \max\left(\hat{d}_{ik\xi}(t-1) - \bar{d}_i \sqrt{\frac{\log_+\left(\frac{TK}{N_{\xi,k}(t-1)|\mathcal{X}|}\right)}{N_{\xi,k}(t-1)}}, 0\right),$$

where $\log_+(x) = \log(x) \mathbf{1}_{\{x \ge 1\}}$.

Since $U_{jk\xi}(t) \ge L_{jk\xi}(t)$, by Eq (EC.31) we have

$$\begin{aligned}
(\ddagger) &\le \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(b_{jk\xi(t)} - U_{jk\xi(t)}(t) + U_{jk\xi(t)}(t) - L_{jk\xi(t)}(t) + L_{jk\xi(t)}(t) - b_{jk}\xi(t))^+ x_{\xi(t),k}(t)\right] \\
&\le \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(b_{jk\xi(t)} - U_{jk\xi(t)}(t))^+ x_{\xi(t),k}(t)\right] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{E}\left[(U_{jk\xi(t)}(t) - L_{jk\xi(t)}(t)) x_{\xi(t),k}(t)\right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(L_{jk\xi(t)}(t)-b_{jk\xi(t)}(t))^{+}x_{\xi(t),k}(t)\right]\\
\le&\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(b_{jk\xi(t)}-U_{jk\xi(t)}(t))^{+}\right]+\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(U_{jk\xi(t)}(t)-L_{jk\xi(t)}(t))x_{\xi(t),k}(t)\right]\\
&+\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(L_{jk\xi(t)}(t)-b_{jk\xi(t)}(t))^{+}\right].
\end{aligned}
$$

In the second step, we again use the fact that $(\alpha+\beta+\gamma)^{+}\le\alpha^{+}+\beta^{+}+\gamma^{+}$ for any $\alpha,\beta,\gamma\in\mathbb{R}$. In the last step, we use the fact that $0\le x_{\xi(t),k}(t)\le 1$.

By Lemma EC.3, for any $\xi\in\mathcal{X}$, we have

$$
\sum_{k=1}^{K}\sum_{t=1}^{T}\mathrm{E}\left[(b_{jk\xi}-U_{jk\xi}(t))^{+}\right]\le(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot 6\sqrt{KT|\mathcal{X}|}.
$$

Since $\xi(t)$ is independent of θ and $\mathcal{H}_{t-1}$, we have

$$
\mathrm{E}[(b_{jk\xi}-U_{jk\xi}(t))^{+}\mid\xi(t)=\xi]=\mathrm{E}[(b_{jk\xi}-U_{jk\xi}(t))^{+}].
$$

Therefore,

$$
\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(b_{jk\xi(t)}-U_{jk\xi(t)}(t))^{+}\right]\le(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot 6\sqrt{KT|\mathcal{X}|}.
$$

Similarly, we have

$$
\sum_{k=1}^{K}\sum_{t=1}^{T}\mathrm{E}\left[(L_{jk\xi(t)}(t)-b_{jk\xi(t)})^{+}\right]\le(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot 6\sqrt{KT|\mathcal{X}|}.
$$

Because θ and $\theta(t)$ are identically distributed given $\mathcal{H}_{t-1}$ and are independent of $\xi(t)$, we have $\mathbb{P}(\theta\mid\mathcal{H}_{t-1},\xi(t))=\mathbb{P}(\theta(t)\mid\mathcal{H}_{t-1},\xi(t))$, and thus $\mathbb{P}(b_{jk\xi(t)}\mid\mathcal{H}_{t-1},\xi(t))=\mathbb{P}(b_{jk\xi(t)}(t)\mid\mathcal{H}_{t-1},\xi(t))$. Since $L_{jk\xi}(t)$ is deterministic given $\mathcal{H}_{t-1}$, using the law of iterated expectation, we have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{t=1}^{T}\mathrm{E}\left[(L_{jk\xi(t)}(t)-b_{jk\xi(t)}(t))^{+}\right]=&\sum_{k=1}^{K}\sum_{t=1}^{T}\mathrm{E}\left[\mathrm{E}[(L_{jk\xi(t)}(t)-b_{jk\xi(t)}(t))^{+}\mid\mathcal{H}_{t-1},\xi(t)]\right]\\
=&\sum_{k=1}^{K}\sum_{t=1}^{T}\mathrm{E}\left[\mathrm{E}[(L_{jk\xi(t)}(t)-b_{jk\xi(t)})^{+}\mid\mathcal{H}_{t-1},\xi(t)]\right]\\
=&\sum_{k=1}^{K}\sum_{t=1}^{T}\mathrm{E}\left[(L_{jk\xi(t)}(t)-b_{jk\xi(t)})^{+}\right]\\
\le&(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot 6\sqrt{KT|\mathcal{X}|}.
\end{aligned}
$$

In addition, by Lemma EC.4, we have

$$
\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(U_{jk\xi(t)}(t)-L_{jk\xi(t)}(t))x_{\xi(t),k}(t)\right]\le(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot 24\sqrt{|\mathcal{X}|KT\log K}.
$$

Combining the above results, we have

$$
(\ddagger)\le(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot(6+6+24\sqrt{\log K})\sqrt{|\mathcal{X}|KT}\le(\sum_{i=1}^{N}a_{ij}\bar{d}_i)\cdot 36\sqrt{|\mathcal{X}|KT\log K}. \tag{EC.32}
$$

Adding Eq (EC.29), (EC.30) and (EC.32), we complete the proof of the following inequality:

$$\mathrm{E}\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} a_{ij}D_i(t) - I_j\right)^+\right] \le (\sum_{i=1}^{N} a_{ij}\bar{d}_i)\cdot 38\sqrt{KT\log K}, \quad \forall j=1,\ldots,M. \tag{EC.33}$$

In the final step, combining Eq (EC.28) and (EC.33), we establish an upper bound of the Bayesian regret of TS-contextual as

$$\mathrm{BayesRegret}(T) \le \left(18p_{\max} + 38\left(\sum_{j=1}^{M} p^j_{\max}\sum_{i=1}^{N} a_{ij}\bar{d}_i\right)\right)\cdot\sqrt{|\mathcal{X}|TK\log K}.$$

EC.5. Proof for TS-BwK

THEOREM 5. *The Bayesian regret of TS-BwK is bounded by*

$$\mathrm{BayesRegret}(T) \le \left(\frac{37M}{c_{\min}}\log(T) + 20\right)\sqrt{KT\log K},$$

where $c_{\min} = \min_{j\in[M]} c_j$.

Proof. The total expected reward of TS-BwK is bounded by

$$\mathrm{E}[\mathrm{Reward}(T)] \ge \mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k x_k(t)\right] - \sum_{t=1}^{T}\mathbb{P}\left(\sum_{s=1}^{t} B_j(t) > I_j, \forall j\in[M]\right). \tag{EC.34}$$

Note that $\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k x_k(t)\right]$ is the total expected reward if we assume the process ends at period T, while $\mathbb{P}\left(\sum_{s=1}^{t} B_j(t) > I_j, \forall j\in[M]\right)$ is the probability that some resource is depleted and the process has ended before period t. Since we assume that reward in each period is in $[0,1]$, we obtain Eq (EC.34).

Let x^* be the optimal solution of the LP defined in TS-BwK if $r_k(t) = r_k, b_{jk}(t) = b_{jk}, \forall k\in[K], j\in[M]$. In Section 3.1.1, we have shown that

$$\mathrm{BayesRegret}(T) \le T\cdot\mathrm{E}[\sum_{k=1}^{K} r_k x_k^*] - \mathrm{E}[\mathrm{Reward}(T)].$$

Therefore, by Eq (EC.34), we have

$$\mathrm{BayesRegret}(T) \le \underbrace{\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k(x_k^* - x_k(t))\right]}_{\text{(I)}} + \underbrace{\sum_{t=1}^{T}\mathbb{P}\left(\sum_{s=1}^{t} B_j(t) > I_j, \forall j\in[M]\right)}_{\text{(II)}}. \tag{EC.35}$$

To complete the proof, we show that both (I) and (II) are bounded. For term (I), we have

$$\mathrm{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} r_k(x_k^* - x_k(t))\right] \le 18\sqrt{TK\log K}. \tag{EC.36}$$

This is an immediate result from part (I) of the proof for TS-fixed.

To bound term (II), we use Markov's inequality. By definition, $I_j = c_j T$, so for any $t\in[T]$

$$\mathbb{P}\left(\sum_{s=1}^{t} B_j(s) > I_j\right) \le \mathbb{P}\left(\left(\sum_{s=1}^{t} B_j(s) - c_j t\right) > (T-t)c_j\right) \le \frac{\mathrm{E}\left[\left(\sum_{s=1}^{t} B_j(s) - c_j t\right)^+\right]}{(T-t)c_j}.$$

Using part (II) of the proof for TS-fixed by replacing T with any integer $t\in[T]$, we get

$$\mathrm{E}\left[\left(\sum_{s=1}^{t} B_j(s) - c_j t\right)^+\right] \le 37\sqrt{tK\log K}.$$

Therefore,

$$\mathbb{P}\left(\sum_{s=1}^{t} B_j(s) > I_j\right) \le \frac{37\sqrt{tK\log K}}{(T-t)c_j}. \tag{EC.37}$$

Let $c_{\min} = \min_{j\in[M]} c_j$. By Eq (EC.37), we get

$$\begin{aligned}
\text{(II)} \le & \sum_{t=1}^{T} \mathbb{P}\left(\sum_{s=1}^{t} B_j(t) > I_j, \forall j \in [M]\right) \\
= & \sum_{t=1}^{T-\lfloor\sqrt{T}\rfloor-1} \mathbb{P}\left(\sum_{s=1}^{t} B_j(t) > I_j, \forall j \in [M]\right) + \sum_{t=T-\lfloor\sqrt{T}\rfloor}^{T} \mathbb{P}\left(\sum_{s=1}^{t} B_j(t) > I_j, \forall j \in [M]\right) \\
\le & \sum_{t=1}^{T-\lfloor\sqrt{T}\rfloor-1} \sum_{j=1}^{M} \mathbb{P}\left(\sum_{s=1}^{t} B_j(t) > I_j\right) + \lfloor\sqrt{T}\rfloor + 1 \\
\le & \sum_{t=1}^{T-\lfloor\sqrt{T}\rfloor-1} \sum_{j=1}^{M} \frac{37\sqrt{tK\log K}}{(T-t)c_j} + \sqrt{T} + 1 \\
\le & \sum_{j=1}^{M} \int_{0}^{T-\lfloor\sqrt{T}\rfloor} \frac{37\sqrt{tK\log K}}{(T-t)c_j} + \sqrt{T} + 1 \\
\le & \frac{37M}{c_{\min}} \sqrt{TK\log K} \cdot \log(T) + \sqrt{T} + 1.
\end{aligned}$$

In the last step, we use the fact that for $T \ge 1$,

$$\int_0^{T-\sqrt{T}} \frac{\sqrt{t}}{T-t} dt = 2\sqrt{T}\log\left(T^{1/4} + \sqrt{\sqrt{T}-1}\right) - 2\sqrt{T-\sqrt{T}} \le \sqrt{T}\log(T).$$

In sum, combining the bounds for both terms in Eq (EC.35), we have

$$\text{BayesRegret}(T) \le 18\sqrt{KT\log K} + \frac{37M}{c_{\min}}\sqrt{KT\log K}\log(T) + \sqrt{T} + 1 \le \left(\frac{37M}{c_{\min}}\log(T) + 20\right)\sqrt{KT\log K}.$$

EC.6. Lemmas

We prove a few lemmas based on the result of Bubeck and Liu (2013), Theorem 1. To simplify notation, we assume $D_i(t) \in [0,1]$ for all $i \in [N]$. In the general case where $D_i(t) \in [0, \bar{d}_i]$, all the bounds in the lemmas are scaled by a constant factor of $\bar{d}_i$.

EC.6.1. Discrete Price Setting

We first consider the model setting in Section 2.1. Let d_{ik} be the mean demand of product i under price vector p_k. Let $N_k(t-1)$ be the number of times that price vector p_k is offered over the first $t-1$ periods, and let $\hat{d}_{ik}(t-1) = (\sum_{s=1}^{t-1} \mathbf{1}_{\{P(s)=p_k\}} D_i(s))/N_k(t-1)$ be the average demand of product i when price vector p_k is offered in the first $t-1$ periods. Define the following upper confidence bound (UCB) function

$$U_{ik}(t) = \min\{1, \hat{d}_{ik}(t-1) + \sqrt{\frac{\log\left(\frac{TK}{N_k(t-1)}\right)}{N_k(t-1)}}\},$$

and the lower confidence bound (LCB) function

$$L_{ik}(t) = \max\{0, \hat{d}_{ik}(t-1) - \sqrt{\frac{\log\left(\frac{TK}{N_k(t-1)}\right)}{N_k(t-1)}}\}.$$

LEMMA EC.1. *For any $k \in [K]$ and $t \in [T]$, we have*

$$\mathrm{E}\left[(d_{ik} - U_{ik}(t))^+\right] \le 6\frac{1}{\sqrt{KT}}, \quad \textit{and} \quad \mathrm{E}\left[(L_{ik}(t) - d_{ik})^+\right] \le 6\frac{1}{\sqrt{KT}}.$$

Proof. This is a special case of Lemma EC.3 when $|\mathcal{X}| = 1$. □

LEMMA EC.2. *It holds that*

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(U_{ik}(t) - d_{ik})x_k(t)\right] \le 12\sqrt{KT\log K}, \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(d_{ik} - L_{ik}(t))x_k(t)\right] \le 12\sqrt{KT\log K}.$$

Proof. This is a special case of Lemma EC.4 when $|\mathcal{X}| = 1$. □

EC.6.2. Contextual Setting

We then consider the contextual setting in Section 4.2. Let $d_{ik\xi} := d_{ik}(\xi \mid \theta)$ be the mean demand of product i under price vector p_k and context ξ. Let $N_{\xi,k}(t-1)$ be the number of times that price vector p_k is offered with context ξ over the first $t-1$ periods, and let $\hat{d}_{ik\xi}(t-1) = (\sum_{s=1}^{t-1}\mathbf{1}_{\{P(s)=p_k,\xi(s)=\xi\}}D_i(s))/N_{\xi,k}(t-1)$ be the average demand of product i when price vector p_k is offered with context ξ in the first $t-1$ periods. Define the following upper confidence bound (UCB) function

$$U_{ik\xi}(t) = \min\{1, \hat{d}_{ik\xi}(t-1) + \sqrt{\frac{\log_+\left(\frac{TK}{N_{\xi,k}(t-1)|\mathcal{X}|}\right)}{N_{\xi,k}(t-1)}}\},$$

and the lower confidence bound (LCB) function

$$L_{ik}(t) = \max\{0, \hat{d}_{ik}(t-1) - \sqrt{\frac{\log_+\left(\frac{TK}{N_{\xi,k}(t-1)|\mathcal{X}|}\right)}{N_{\xi,k}(t-1)}}\},$$

where $\log_+(x) = \log(x)\mathbf{1}_{\{x\ge 1\}}$.

LEMMA EC.3. *For any $k \in [K]$, $t \in [T]$ and $\xi \in \mathcal{X}$, we have*

$$\mathrm{E}\left[(d_{ik\xi} - U_{ik\xi}(t))^+\right] \le 6\sqrt{\frac{|\mathcal{X}|}{KT}}, \quad \textit{and} \quad \mathrm{E}\left[(L_{ik\xi}(t) - d_{ik\xi})^+\right] \le 6\sqrt{\frac{|\mathcal{X}|}{KT}}.$$

Proof. We consider a multi-armed bandit problem with $|\mathcal{X}| \times K$ arms. The mean reward of arm (ξ, k) is $d_{ik\xi}$. Bubeck and Liu (2013) (Theorem 1, Step 2) shows that if we choose an UCB function as

$$\tilde{U}_{ik\xi}(t) = \hat{d}_{ik\xi}(t-1) + \sqrt{\frac{\log_+\left(\frac{n}{N_{\xi,k}(t-1)K|\mathcal{X}|}\right)}{N_{\xi,k}(t-1)}},$$

where $\log_+(x) = \log(x)\mathbf{1}_{\{x\ge 1\}}$, then

$$\mathrm{E}\left[(d_{ik\xi} - \tilde{U}_{ik\xi}(t))^+\right] \le 6\sqrt{\frac{K|\mathcal{X}|}{n}}, \quad \forall(\xi,k) \in \mathcal{X}\times[K], t \ge 1.$$

The first inequality in the lemma immediately follows by choosing $n = TK^2$. A similar proof can be used to show that $\mathrm{E}\left[(L_{ik\xi}(t) - d_{ik\xi})^+\right] \le 6\sqrt{|\mathcal{X}|}/\sqrt{KT}$. □

LEMMA EC.4. *Let $\xi(t)$ be the context in period t. It holds that*

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(U_{ik\xi(t)}(t)-d_{ik\xi(t)})x_{\xi(t),k}(t)\right]\le 12\sqrt{|\mathcal{X}|KT\log(K)},$$

and

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(d_{ik\xi(t)}-L_{ik\xi(t)}(t))x_{\xi(t),k}(t)\right]\le 12\sqrt{|\mathcal{X}|KT\log(K)}.$$

Proof. Let $k(t)\in[K]$ be the index of the price vector chosen at period t. Since $k(t)=k$ is chosen independently with probability $x_{\xi(t),k}(t)$, we have

$$\begin{aligned}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(U_{ik\xi(t)}(t)-d_{ik\xi(t)})x_{\xi(t),k}(t)\right]&=\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(U_{ik\xi(t)}(t)-d_{ik\xi(t)})\mathbf{1}_{\{k(t)=k\}}\right]\\&=\sum_{t=1}^{T}\mathrm{E}\left[(U_{ik(t)\xi(t)}(t)-d_{ik(t)\xi(t)})\right].\end{aligned}$$

Let $K'=|\mathcal{X}|K$. If $K'\ge T$, the result holds trivially since

$$\sum_{t=1}^{T}\mathrm{E}\left[U_{ik(t)\xi(t)}(t)-d_{ik(t)\xi(t)}\right]\le T\cdot 1\le\sqrt{K'T}.$$

If $K'\le T$, we consider a multi-armed bandit problem with $|\mathcal{X}|\times K$ arms. The mean reward of arm (ξ,k) is $d_{ik\xi}$. Bubeck and Liu (2013) (Theorem 1, Step 3) shows that if we choose an UCB function as

$$\tilde{U}_{ik\xi}(t)=\hat{d}_{ik\xi}(t-1)+\sqrt{\frac{\log_+\left(\frac{n}{N_{\xi,k}(t-1)K|\mathcal{X}|}\right)}{N_{\xi,k}(t-1)}},$$

where $\log_+(x)=\log(x)\mathbf{1}_{\{x\ge 1\}}$, then

$$\sum_{t=1}^{n}\mathrm{E}\left[\tilde{U}_{ik(t)\xi(t)}(t)-d_{ik(t)\xi(t)}\right]\le 8\sqrt{nK'}.$$

We select $n=T$. Note that

$$U_{ik\xi}(t)\le\min\left(\hat{d}_{ik\xi}(t-1)+\sqrt{\frac{\log_+\left(\frac{T}{N_{\xi,k}(t-1)K|\mathcal{X}|}\right)+\log(K^2)}{N_{\xi,k}(t-1)}},1\right)\le\tilde{U}_{ik\xi}(t)+\min\left(\sqrt{\frac{2\log(K)}{N_{\xi,k}(t-1)}},1\right),$$

so

$$\sum_{t=1}^{T}\mathrm{E}\left[U_{ik(t)\xi(t)}(t)-d_{ik(t)\xi(t)}\right]\le 8\sqrt{TK'}+\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{\xi\in\mathcal{X}}\mathrm{E}\left[\min\left(\sqrt{\frac{2\log(K)}{N_{\xi,k}(t-1)}},1\right)\mathbf{1}_{\{k(t)=k,\xi(t)=\xi\}}\right].$$

Russo and Van Roy (2014) (Proposition 1) shows that

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{\xi\in\mathcal{X}}\mathrm{E}\left[\min\left(\sqrt{\frac{2\log(K)}{N_{\xi,k}(t-1)}},1\right)\mathbf{1}_{\{k(t)=k,\xi(t)=\xi\}}\right]\le K'+2\sqrt{2\log(K)K'T}\le 4\sqrt{\log(K)K'T}.$$

In sum, we have

$$\sum_{t=1}^{T}\mathrm{E}\left[U_{ik(t)\xi(t)}(t)-d_{ik(t)\xi(t)}\right]\le 12\sqrt{K'T\log(K)}=12\sqrt{|\mathcal{X}|KT\log(K)}.$$

Similarly, one can show that

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{E}\left[(d_{ik\xi(t)}-L_{ik\xi(t)}(t))x_{\xi(t),k}(t)\right]=\sum_{t=1}^{T}\mathrm{E}\left[d_{ik(t)\xi(t)}-L_{ik(t)\xi(t)}(t)\right]\le 12\sqrt{|\mathcal{X}|KT\log K}.\quad\square$$